Development and application of new statistical methods for the analysis of multiple phenotypes to investigate genetic associations with cardiometabolic traits

Dissertation

zur Erlangung des akademischen Grades
doctor rerum naturalium (Dr. rer. nat.)
im Fach Informatik

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin

von

Dipl.-Math. M.Sc. Stefan Konigorski

Präsidentin der Humboldt-Universität zu Berlin: Prof. Dr.-Ing. Dr. Sabine Kunst
Dekan der Mathematisch-Naturwissenschaftlichen Fakultät: Prof. Dr. Elmar Kulke

Gutachter/innen:
1. Prof. Dr. Marius Kloft
2. Prof. Dr. Tobias Pischon
3. Prof. Dr. Yildiz E. Yilmaz

Tag der mündlichen Prüfung: 15. März 2018

Zusammenfassung

Die letzten Jahre waren durch weitreichende biotechnologische Weiterentwicklungen gekennzeichnet, welche die Untersuchung der Zusammenhänge von genetischen und molekularen Markern mit komplexen Krankheiten in bis dato nicht bekannter Tiefe ermöglichen. Darüberhinaus werden oft multiple Traits von Phänotypen erhoben und liegen zur Analyse vor. Um die Information dieser reichen und komplexen Daten jedoch nutzen zu können um biologische Erkenntnisse zu gewinnen, sind angemessene statistische Methoden notwendig. In der vorliegenden Dissertation liegt der Fokus auf Genassoziationsstudien mit multiplen Phänotypen, insbesondere auf der Analyse von (i) seltenen genetischen Varianten und (ii) kausalen direkten genetischen Effekten. Für solche Untersuchungen sind vorhandene statistische Methoden nicht immer valide.

Mit diesem Hintergrund ist das erste Ziel dieser Arbeit, zwei neue statistische Methoden für Assoziationsanalysen von genetischen Markern mit multiplen Phänotypen zu entwickeln, die Methoden effizient und robust zu implementieren, so dass sie für die Analyse von Daten aus Hochdurchsatzverfahren benutzt werden können, und die Methoden im Vergleich zu anderen statistischen Ansätzen in realistischen Szenarien zu überprüfen. Der erste Ansatz, genannt *C-JAMP (Copula-based Joint Analysis of Multiple Phenotypes)*, ermöglicht multiple Traits gemeinsam, bedingt auf genetische Varianten und Kovariaten, zu modellieren um die statistische Power von Assoziationstests der genetischen Varianten zu erhöhen. Der zweite Ansatz, genannt *CIEE (Causal Inference using Estimating Equations)*, ermöglicht den indirekten genetischen Effekt über intermediäre Phänotypen zu entfernen um direkte genetische Effekte valide und robust zu schätzen und testen. C-JAMP wird in dieser Arbeit für Genassoziationsstudien von seltenen genetischen Varianten mit quantitativen Traits evaluiert, und CIEE für Genassoziationsstudien von häufigen genetischen Varianten mit quantitativen Traits und Ereigniszeiten als primären Phänotypen.

Die Ergebnisse von umfangreichen Simulationsstudien zeigen, dass beide Methoden eine unverzerrte und effiziente Parameterschätzung ermöglichen und die statistische Power von Assoziationstests im Vergleich zu existierenden Methoden erhöhen können - welche ihrerseits in vielen Szenarien keine valide Analyse erlauben. Diese Eigenschaften von C-JAMP und CIEE werden in Anwendungen für das zweite Ziel dieser Arbeit genutzt, welches darin besteht, neue genetische und transkriptomische Marker für kardiometabolische Traits zu identifizieren. In der ersten Anwendung wird ein existierender empirischer Datensatz analysiert, welcher genetische Marker, Genexpressionsmaße, und Blutdruckmaße enthält, um neue Blutdruckmarker zu identifizieren. In einer zweiten Anwendung werden Daten einer Studie analysiert, die im Rahmen der Dissertation durchgeführt wurde, um den Zusammenhang von genetischen Markern und Genexpressionsmaßen aus RNA-Sequenzierung mit Adipositasmaßen aus Magnetresonanztomographiescans zu untersuchen.

In den Analysen konnten mehrere neue Kandidatenmarker und -gene für Blutdruck und Adipositas identifiziert werden. Dies unterstreicht den Wert, neue statistische Methoden zu entwickeln, evaluieren, und implementieren - um neue Erkenntnisse über Krankheitsmarker zu gewinnen und Hypothesen über zugrundeliegende biologische Prozesse für Follow-up Studien zu generieren. Für beide entwickelten Methoden sind R Pakete verfügbar, die ihre Anwendung in weiteren empirischen Studien ermöglichen.

Summary

In recent years, the biotechnological advancements have allowed to investigate associations of genetic and molecular markers with complex diseases in much greater depth. In addition, multiple phenotypic measures are often available for the analysis. However, in order to make use of the rich and complex data and to provide valid insights, appropriate statistical methods are needed for the investigation of biologically meaningful models. In this thesis, the focus is on genetic association studies with multiple phenotypes, and in particular on (i) the analysis of rare genetic variants and (ii) the identification of direct genetic effects. For such studies, available statistical methods are not always valid.

With this background, the first aim of this thesis is to develop two new statistical methods for the association analysis of genetic markers with multiple phenotypes, to implement them in a computationally efficient and robust manner so that they can be used for large-scale analyses, and evaluate them in comparison to existing statistical approaches under realistic scenarios. The first approach, called the *copula-based joint analysis of multiple phenotypes (C-JAMP)* method, allows investigating multiple traits in a joint model conditional on genetic variants and covariates to improve the statistical power for identifying associated genetic variants, and is evaluated for genetic association analyses of rare genetic variants with quantitative traits. The second proposed approach, called the *causal inference using estimating equations (CIEE)* method, allows removing indirect genetic effects through intermediate phenotypes in order to estimate and test direct genetic effects on the primary phenotype in a valid and robust manner, and is evaluated for genetic association analyses of common genetic variants with quantitative and time-to-event primary phenotypes.

The results of extensive simulation studies show that both approaches provide an unbiased and efficient estimation of parameters and can improve the power of association tests in comparison to existing approaches, which are invalid in many scenarios. These properties of C-JAMP and CIEE are used in application studies for the second goal of this thesis, to identify novel genetic and transcriptomic candidate markers associated with cardiometabolic traits. C-JAMP and CIEE are applied to one existing empirical dataset which contains genetic markers, gene expression measures and blood pressure traits, to identify novel blood pressure loci. Furthermore, they are applied in an empirical study which was planned and conducted as part of this dissertation, to assess genetic markers and gene expression from RNA-sequencing and test their association with obesity phenotypes from magnetic resonance imaging scans.

In the analyses, several novel candidate markers and genes are identified, which highlights the merit of developing, evaluating, and implementing appropriate statistical approaches in order to gain new insights about disease markers and to build hypotheses about biological processes for follow-up studies. R packages are available for both methods and enable their application in further empirical studies.

Acknowledgements

I am very thankful for the opportunity to work on the projects involved in this dissertation, which has been supported and made possible by the contribution of many different people.

First, I want to express my deepest gratitude to my supervisors Prof. Dr. Tobias Pischon, Prof. Dr. Yildiz Yilmaz, and Prof. Dr. Marius Kloft for their guidance and support. Before coming to Prof. Pischon's group at the Max Delbrück Center (MDC) in Berlin, I was hoping to be more involved in the epidemiological study design and in the generation of molecular data. I thank him for making this possible, through which I learned immensely about molecular epidemiology, sequencing as well as traditional technologies, but also for the freedom and support to work on the development of statistical methods! In addition, he was always supportive and made it possible to present my work at different conferences, to be involved in teaching activities, and also supported my research stay at the Memorial University of Newfoundland with Prof. Yilmaz. Professor Yilmaz introduced me first to the concept of copula functions during my Master studies at the University of Toronto, as well as to estimating equations during my visit in St. John's, and helped me to get started on these projects. I appreciate the chance to work on these methodological projects which are the foundation of this thesis, the many hours that she devoted to explain the methodologies, discuss study concepts, and her dedicated supervision of the ongoing work. Plus, the great times when she and her husband Prof. Dr. Candemir Cigsar showed me around in Toronto, St. John's, and invited me to the Derwish dance and the cave restaurants in Cappadocia will always be fond memories! Finally, I am very grateful that Prof. Kloft accepted me as his PhD student at Humboldt-University of Berlin and gave me a chance to learn from his expertise. He helped me, always smiling, with my PhD plans and all time restrictions, and it doesn't seem like a coincidence when I think back of our first meeting years ago over some Flammkuchen! I enjoyed very much and learned a lot from our meetings and discussions about the methodologies, often taking a step back to look at the bigger methodological concepts and putting them into perspective, and linking them to computational and machine learning concepts. Further, I take away a renewed motivation to work on machine learning concepts which may now be more relevant for biomedical studies than ever. Thank you all for being such great mentors and more, I am looking forward to many more interesting discussions and collaborations with you!

In addition to my supervisors, the comments and suggestions from Prof. Dr. Norbert Hübner and Prof. Dr. Heiner Boeing during the committee meetings helped me a lot for progressing with and completing my thesis. During the work on my thesis, I had a great time at the MDC and with everyone in our Molecular Epidemiology group, who welcomed me

from the first day. I thank Dr. Jürgen Janke and Dr. Giannino Patone for their patience and teaching me about PCR, ELISA, RNA-Sequencing, the bioinformatic pipelines, and giving me the chance to ask all these basic questions! Similarly, I had a great time during the 3 months in St. John's, getting to know the land and people, and learning much more about statistics and genetics. Thank you for welcoming me Prof. Yilmaz and Prof. Cigsar, and everyone else at MUN, and to Prof. Yilmaz, Prof. Pischon, and the MDC for making it possible!

Also more general, I thank Prof. Pischon, Prof. Yilmaz, the MDC, and the Metabolic Dysfunction initiative of the Helmholtz Association for their funding and financial support that made my work and the underlying experiments possible. The analysis of the Genetic Analysis Workshop 19 data in the first data application was made possible through the Genetic Analysis Workshops with a big help by Vanessa Olmo, Prof. Dr. Laura Almasy, and Dr. Jean MacCluer. For the second application in the obesity study, a large number of individuals were involved at all stages. I thank the participants of the EPIC study, Prof. Boeing and Dr. Manuela Bergmann for their overall coordination, approval of the analysis plans and availability of the data and biomaterial, Henning Damm, Sarah Moreno Garcia, Dr. Kathrin Saar, Susanne Blachut, and the Genomics Sequencing platform members at the MDC for the molecular data preparation and performing the experiments, and all other groups and members that I forgot to mention here.

Most of all, I want to express special thanks to my wife Ga Young for her continuous and everyday love and support through all rainy and happy days, Olmang, and my family!

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AFT	Accelerated failure time
aSPU	Adaptive sum of powered score (tests)
AT	Adipose tissue
BH	Benjamini-Hochberg
BMI	Body mass index
BP	Blood pressure
CAT	Coronary adipose tissue
CCA	Canonical correlation analysis
CDF	Cumulative distribution function
C-JAMP	Copula-based joint analysis of multiple phenotypes
CIEE	Causal inference using estimating equations
DAG	Directed acyclic graph
ELISA	Enzyme-linked immunosorbent assay
EPIC	European Prospective Investigation into Cancer and Nutrition
eQTL	Expression quantitative trait loci
GAW	Genetic Analysis Workshop
GC (content)	Guanine-cytosine (content)
GE	Gene expression
GEE	Generalized estimating equation
GWAS	Genome-wide association study
HC	Hip circumference
KS	Kolmogorov-Smirnov
LD	Linkage disequilibrium
LM	Linear model
LRT	Likelihood ratio test
LS	Least squares
MAC	Minor allele count
MAD	Median absolute deviation
MAF	Minor allele frequency
MANOVA	Multivariate analysis of variance
ML	Maximum likelihood
MLE	Maximum likelihood estimate
MMT	Multi-marker test
MR	Multiple regression
MRI	Magnetic resonance imaging

mRNA	Messenger ribonucleic acid
MURAT	Multivariate rare-variant association test
MVN	Multivariate normal distribution
PCR	Polymerase chain reaction
PDF	Probability density function
PH	Proportional hazard
Q-Q plot	Quantile-quantile plot
RNA-Seq	RNA-sequencing
RR	Regression of residuals
SAT	Subcutaneous adipose tissue
SEM	Structural equation modeling
SKAT	Sequence kernel association test
SKAT-O	Optimal sequence kernel association test
SMT	Single-marker test
SNV	Single nucleotide variant
SBP	Systolic blood pressure
TAT	Total adipose tissue
VAT	Visceral adipose tissue
WC	Waist circumference
WHR	Waist-to-hip-ratio

Notation

$X, Y, ...$	(Random) variables
$\boldsymbol{X}, \boldsymbol{Y}, ...$	Variable vectors (or matrices)
$x, y, ...$	Observed realizations of X, Y (i.e., known constants)
$\alpha, \beta, ...$	Single parameters
$\boldsymbol{\alpha}, \boldsymbol{\beta}, ...$	Parameter vectors
i	Individual, $i = 1, ..., n$
X_j	Single nucleotive variant (SNV) j, $j = 1, ..., k$
Y_l	(Primary) phenoype of interest l, $l = 1, ..., p$
G_s	Gene expression level of gene s, $s = 1, ..., d$
Z	Covariate
K	Secondary phenotype
L	Measured confounding variable of the effect of K on Y
U	Unmeasured confounding variable of the effect of K on Y
α_{XY}, β_{XY}	Effect of the genetic marker X on the phenotype Y
ψ, φ, θ	Dependence parameters of copula functions
ϕ	Probability density function of the standard normal distribution
Φ	Cumulative distribution function of the standard normal distribution

Chapter 1

Introduction

1.1 Motivation and aim

Current trends in society feature aspects of health and technology across many different disciplines, lifestyles, and areas. In everyday life, smartphone apps can be used to measure physiological data including body temperature, heart rate, or physical activity, to count calories intake or also blood glucose levels. In the healthcare system, efforts are made to digitalize and integrate patient registries, and to combine electronic health records. Finally in biomedical research, the technological advancements allow to study molecular processes in greater depth, collaborative efforts are made to enable the analysis of large multi-center cohorts, and efficient bioinformatics pipelines and quality control recommendations are set up. All these examples create and deal with data that can be used to gain a deeper understanding of the contributing risk factors to health and disease. However, due to the resulting complex and high-dimensional datasets, such studies require appropriate and powerful statistical methods for the analysis of the multitude of predictor and outcome variables as well as an efficient computational implementation, in combination with appropriate epidemiological study designs and biomedical expertise.

Regarding the development and application of statistical methods, it is critical that any proposed method is (i) properly evaluated in realistic scenarios and compared to existing methods, (ii) implemented efficiently and in a robust manner so that it can be applied to large-scale data and integrate the information from the multitude of measured molecular and phenotypic variables, and that (iii) any new identified biological signals in the analyses are interpretable and carry useful biological information on new markers or structures. While these points might seem intuitive, they have not always been considered in previous studies. The focus in this thesis is on genetic association analyses and in particular on rare variant analysis and the identification of direct genetic effects. It will be shown that in these fields, some recently proposed methods are invalid for the analysis of empirical data, do not have the proposed statistical properties in a finite-sample setting, and are even inferior to simple standard approaches in many scenarios.

The first aim of this thesis is then to develop two new statistical methods called C-JAMP and CIEE, for the association analysis of genetic markers with multiple outcome variables (i.e., phenotypes), to implement them in a computationally efficient and robust manner so

that they can be used for large-scale analyses, and evaluate them in comparison to standard statistical approaches and recently proposed approaches under realistic scenarios.

Aim 1

Develop, implement, and evaluate two novel powerful statistical approaches for genetic association analyses with multiple phenotypes.

Both methods can be applied to a large variety of datasets, phenotypes, and biological questions, and are presented and evaluated here for specific analyses: The *copula-based analysis of multiple phenotypes* (C-JAMP) method is evaluated for the association analysis of rare genetic variants with multiple quantitative traits, and the *causal inference using estimating equations* (CIEE) method is evaluated for the identification of direct effects of common genetic variants on a primary quantitative or time-to-event phenotype. In an evaluation of existing statistical methods for these analyses, it will be shown that most of them are invalid in some scenarios and the following thesis is proposed:

Thesis

1. **Modeling multiple phenotypes with C-JAMP in a joint model or with CIEE in a directed acyclic graph yields valid and efficient genetic effect estimates, and can increase the power of hypothesis tests compared to separate univariate models.**
2. **The two proposed methods provide an improvement over existing methods for**
   (a) **association analyses of rare genetic variants and**
   (b) **the identification of direct effects of common genetic variants.**

In a second aim, the proposed methods are applied to empirical data in order to identify new candidate genetic markers and genes for cardiometabolic traits. First, genetic associations with blood pressure (BP) are investigated in published data from the Genetic Analysis Workshop (GAW) 19. Second, a study was planned and conducted in order to identify new candidate markers and genes associated with obesity traits, by measuring single nucleotide variants (SNVs) and gene expression levels in adipose tissue (AT) through RNA-sequencing (RNA-Seq; Wang et al., 2009; Ozsolak & Milos, 2010), and applying C-JAMP and CIEE.

Aim 2

Identify novel genetic and transcriptomic effects on cardiometabolic traits by conducting empirical studies and applying the proposed statistical methods.

As a result, the interdisciplinary projects in this thesis incorporate statistical, computational, bioinformatic and molecular epidemiological aspects, proposing appropriate methods and evaluating them in a meaningful way for aim 1, as well as planning and conducting a molecular epidemiological study and using the developed methods in the analysis for aim 2.

1.2 Thesis outline

Regarding the structure of this thesis, the next section 1.3 gives an overview about the scientific contributions in this thesis. Following the introductory chapter, chapter 2 gives some relevant epidemiological, statistical, and computational background. More specifically, section 2.1 outlines the public health relevance of investigating risk factors for cardiometabolic traits and describes their assessment and results from genetic association studies. Section 2.2 gives an overview on existing statistical approaches and methods for the analysis of complex genotype-phenotype data. First, methods to analyze multiple genetic markers jointly are described, which have mainly been developed for the analysis of rare genetic variants. Second, an overview about methods to analyze multiple phenotypes is given. Section 2.3 presents some relevant computational aspects for the implementation of statistical approaches and analysis of high-dimensional genotype-phenotype data.

Next, the development of two novel statistical approaches is presented in chapter 3, including their theoretical derivation and their respective implementation in packages of the statistical software R (R Core Team, 2017). The first proposed approach is called C-JAMP and uses copula functions to build joint models of multiple traits of a phenotype conditional on genetic markers and other covariates (section 3.1). The statistical details of C-JAMP are described in section 3.1.1 with a focus on the analysis of quantitative traits, and section 3.1.2 gives details regarding the implementation of C-JAMP in R. In section 3.2, the second novel approach called CIEE is presented. CIEE fits a directed acyclic graph using estimating equations to estimate and test direct effects of genetic markers on a primary phenotype while removing indirect genetic effects through intermediate phenotypes. The statistical details are described in section 3.2.1 with a focus on the analysis of quantitative and time-to-event primary traits. Implementation details of the developed R package are given in section 3.2.2.

After the development of the new statistical methods, C-JAMP, CIEE and existing approaches are evaluated and compared for association analyses of genetic markers by using simulation studies in chapter 4. First, in section 4.1, existing methods are evaluated and compared for the analysis of rare genetic variants with quantitative phenotypes, to evaluate whether or when popular models aggregating multiple genetic markers in a region (multi-marker tests, MMT: SKAT, Wu et al., 2011; SKAT-O, Lee et al., 2012; and a burden test, Lee et al., 2014) have higher power compared to simple separate Wald-type t-tests of each genetic marker (single-marker tests; SMT). The results show that in all considered scenarios, the SMT allows valid estimation and inference and has a higher power compared to the MMTs in many realistic situations. Consequently, the following evaluation of C-JAMP and CIEE focuses on their use as SMTs. Section 4.2 describes a large simulation study investigating the properties of C-JAMP and compares its performance to the simple SMT based on a univariate model of a trait as well as the univariate MMTs (SKAT, SKAT-O, burden

test), multivariate SMTs (MultiPhen, O'Reilly et al., 2012; aSPU, Kim et al., 2016) and multivariate MMTs (MURAT, Sun et al., 2016b; aSPUset, aSPUset-Score, Kim et al., 2016), for association studies or rare variants. Finally, section 4.3 evaluates the statistical properties of CIEE, and compares its empirical type I error and power to alternative methods for estimating and testing direct effects of common genetic variants, namely traditional regression methods, the structural equation modeling method (Bollen, 1989), and G-estimation methods (Robins, 1986, 1992; Robins & Greenland, 1994; Goetgeluk et al., 2008).

Following the evaluation of the proposed statistical approaches, C-JAMP and CIEE are applied to two different empirical studies in chapter 5. In section 5.1, they are applied in the GAW19 data (Blangero et al., 2016) to infer overall as well as direct genetic effects on BP. The data contains whole genome-sequence data, gene expression in lymphocytes measured with microarrays, BP phenotypes, as well as non-genetic covariates of 81 unrelated individuals from the T2D-GENES Consortium. C-JAMP is used to model the joint distribution of systolic BP (SBP) and gene expression conditional on each SNV within the gene, and conditional on the nongenetic covariates (see Figure 1.1, left panel). CIEE is applied to infer direct genetic effects on SBP while removing indirect genetic effects through gene expression and adjusting for measured and unmeasured confounding effects as illustrated in the underlying directed acyclic graph (DAG) in Figure 1.1 (right panel).

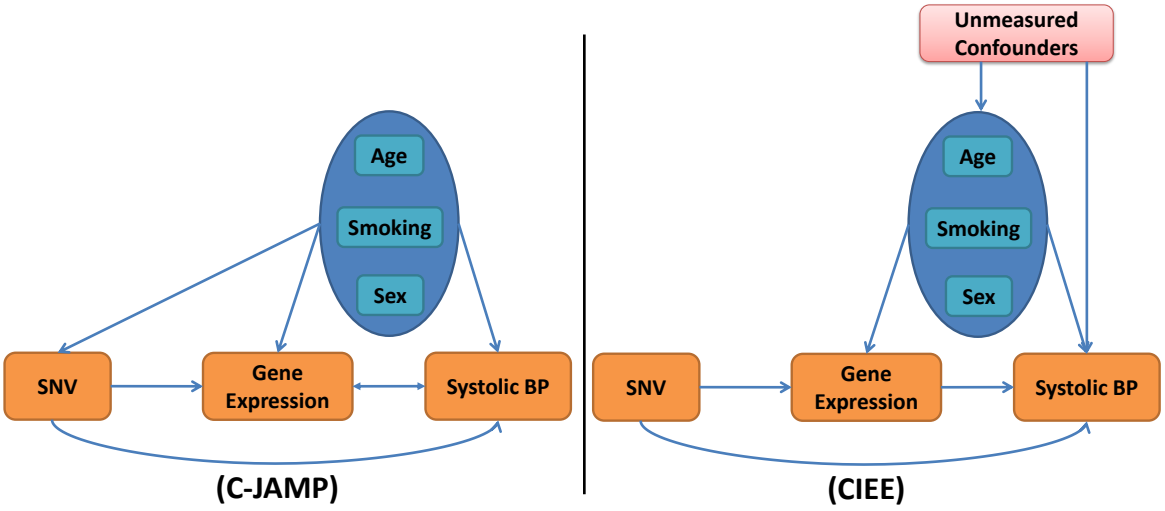


Figure 1.1: Assumed models in the GAW19 analysis, of the potential effects between the SNVs, gene expression levels, systolic BP and non-genetic covariates in the analyses using C-JAMP (left panel) and CIEE (right panel) in section 5.1. In both analyses, systolic BP is considered as a trait censored by anti-hypertensive medication use.

Section 5.2 describes the application of C-JAMP to estimate and test genetic and transcriptomic effects on obesity traits (Figure 1.2, left panel) and of CIEE to estimate and test direct genetic effects on obesity traits while removing indirect genetic effects through gene expression (Figure 1.2, right panel). For the analysis, a study was planned and conducted to assess whole-transcriptome gene expression levels in abdominal AT with RNA-Seq cross-sectionally in a population-based sample of 200 probands from the European Prospective Investigation into Cancer and Nutrition (EPIC) Potsdam substudy (Wientzek et al., 2014; Neamat-Allah et al., 2015). SNVs in coding regions were called from the RNA-Seq data, and

previously validated data from magnetic resonance imaging (MRI) scans gives a direct quantification of body fat mass (measured through the amount of subcutaneous AT, SAT) and body fat distribution (measured through the amount of SAT relative to the amount of total AT, TAT) in different body compartments (Wald et al., 2012; Neamat-Allah et al., 2015).

Finally, chapter 6 summarizes the results from the methods' evaluation, which can be used as a guide for practical data analysis, discusses the two developed methods and their improvement relative to established methods, and discusses the identified new candidate SNVs and genes from the applications in chapter 5. The references for all sections are shown in the following section.

In the appendix, first, a declaration is given that I have worked on this thesis independently in section A.1. Some sections in this thesis including their tables and figures are based on published manuscripts or manuscripts currently under review. A complete list of these manuscripts and publications, an overview about my contributions in the publications, and details regarding which publications are underlying which sections are given in sections A.2-A.3. Further details and results of the analyses in chapters 3-5, supplementary tables and figures, and an outline of the R code of the developed R packages can be found in the further sections A.4-A.7 of the appendix.

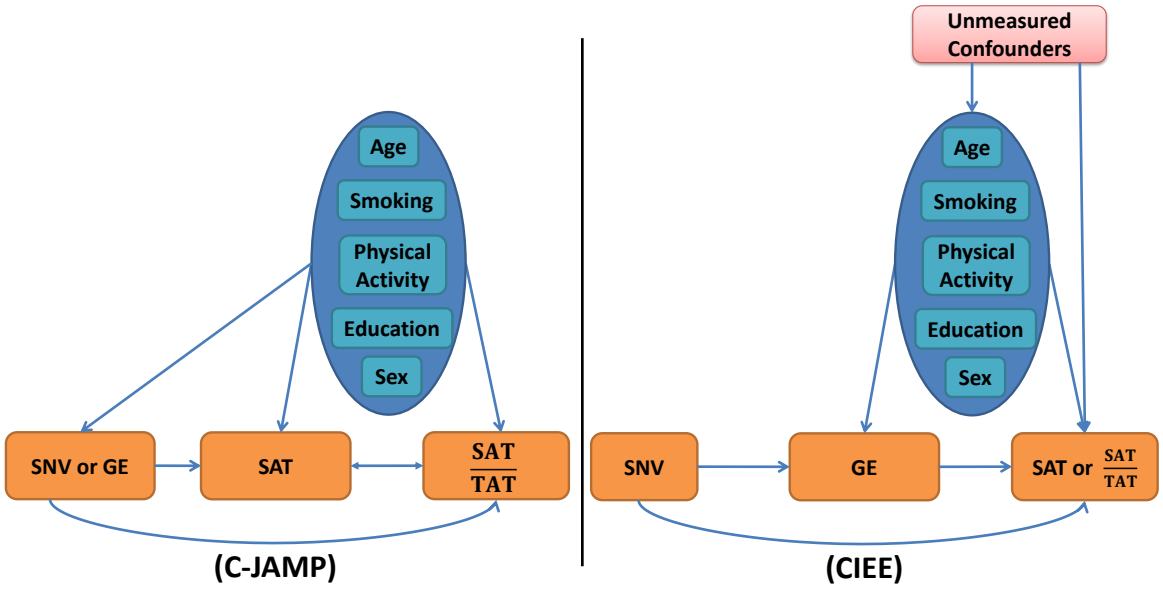


Figure 1.2: Assumed models in the obesity study, of the potential effects between the SNVs, gene expression levels (GE), obesity traits SAT mass and $\frac{\text{SAT}}{\text{TAT}}$ (SAT mass relative to TAT mass) and non-genetic covariates in the analysis of the EPIC Potsdam substudy data using C-JAMP (left panel) and CIEE (right panel) in section 5.2.

1.3 Contributions

In this thesis, two novel statistical approaches for genetic association studies are developed, implemented in R, evaluated in extensive simulation studies, and applied to real data. A detailed overview about my contribution to these developments and the related publications is given in section A.2. To the best of my knowledge, C-JAMP is the first implementation of copula models to model the dependence of multiple phenotypes for genetic association studies

and whose efficient and robust implementation allows performing genome-wide analyses, and CIEE is the first valid approach for the analysis of log-linear models of time-to-event primary phenotypes. Both approaches are evaluated for specific analyses but are widely applicable to different datasets, can adjust for covariates, and can be easily extended to multiple outcomes and adapted to outcomes with different distributions.

The evaluation of existing statistical approaches and of the two proposed approaches in simulation studies yields several new results which are important for practical analyses (and are summarized in chapter 6). Simulation studies were chosen as the primary tool for evaluation since the situations of interest (rare genetic variants, traits of high dependence) investigate parameters on the boundary of the parameter space so that theoretical considerations might not be true, especially in the finite-sample setting. The first new result from the simulation studies is that simple single-marker tests are valid for the analysis of rare variants with quantitative traits and have actually equal or higher power compared to popular multi-marker tests in many realistic scenarios (and lower power in some scenarios). Second, it is shown that a number of existing approaches for rare-variant analyses as well as for the identification of direct genetic effects are invalid, provide biased effect estimates and inflated type I errors. On the other hand, it is shown that both proposed approaches allow unbiased and efficient estimation of overall/direct genetic effects, and can increase the power of genetic association tests of rare variants, and of direct genetic effects compared to existing approaches. Extensive simulation studies show their robustness and efficiency over many different scenarios, and also indicate against which model misspecifications the approaches are robust and against which model specifications they are sensitive which have to be checked in practical applications. In order to conduct these extensive simulation studies, an efficient parallelization of the computations on a computing cluster and integration of the results was necessary to handle the computational burden of more than 300 computing years. For example, for an evaluation of C-JAMP (and the other existing statistical approaches) regarding their empirical type I error for a nominal level of $\alpha = 2.5 \times 10^{-6}$, data was generated and evaluated 10,000,000 times for each null model.

In addition to the evaluation of the existing and new approaches on artificial data, C-JAMP and CIEE were applied to real data in several application studies. For this, one existing dataset was analyzed to identify novel genetic markers for blood pressure, and one empirical study was planned, the biological experiments supervised, the bioinformatic processing and quality control performed, and the different datasets integrated in order to identify novel candidate genetic and transcriptomic markers for obesity. The analyses confirmed the results from the simulation studies that both C-JAMP and CIEE yield smaller p-values compared to univariate approaches and can thereby identify more candidate markers that would be missed by existing approaches. More specifically, the genetic association analyses identified 1 novel candidate SNV for BP, 1 novel candidate SNV for body fat mass and 6 novel candidate SNVs for body fat distribution, and further promising SNVs with suggestive evidence. The transcriptomic association analyses identified 441 genes associated with body fat mass and 225 genes associated with body fat distribution. These results suggest new pathways that might be involved in blood pressure and obesity.

Chapter 2

Background

2.1 Public health relevance and genetic epidemiological background

In this thesis, two novel statistical methods are developed that incorporate multiple phenotypes in genetic association analyses. In the empirical data analysis in chapter 5 using the two novel methods, the focus is on the analysis of cardiometabolic traits, and in particular on the cardiovascular phenotype BP and the metabolic phenotype obesity. Throughout this thesis, the term phenotype is used as a synonym of any measured nongenotypic characteristic, encompassing multiple traits. Cardiometabolic traits in general subsume different cardiovascular and metabolic traits such as abdominal adiposity, hypertension, dyslipidemia, hyperinsulinemia, and glucose intolerance. These clustered disorders, also termed cardiometabolic disease or cardiometabolic syndrome (Castro et al., 2003; Fisher, 2006), are themselves risk factors for some of the most prevalent chronic diseases such as diabetes, heart disease, or stroke. For example, current estimates suggest that the overall global prevalence of hypertension was 26.4% in 2000, and project a prevalence of 29.2% in 2025 equaling 1.56 billion adults (Kearney et al., 2005). Regarding diabetes, current estimates suggest that 415 million adults are affected worldwide with an expected rise to 642 million adults by 2040, equaling a prevalence of 9% and accounting for 12% of global health spendings (Aguiree et al., 2013). The identification of molecular correlates of cardiometabolic traits can contribute to a more detailed understanding of the disease etiology, to an identification of new biomarkers and a deeper understanding of established biomarkers. Such biomarkers can be valuable and informative markers for disease, and can be used in prevention guidelines as well as potential targets for treatment (Tam et al., 2011; Fisman & Tenenbaum, 2014). While the main focus in this thesis will be on genetic association analyses, they are often only a first step in the analysis of complex traits. Assessing gene expression levels can yield more detailed hypotheses about molecular processes and can serve as a starting point to further investigations using high-throughput approaches in the fields of proteomics, metabolomics, epigenetics and others.

As genetic markers, single nucleotide variants (SNVs) are considered throughout this thesis, which are genetic polymorphisms (i.e., variations) of a single nucleotide at specific positions in the genome, with any frequency of the minor and major alleles. The distinction

between common and rare SNVs is usually made based on their minor allele frequency (MAF), and a threshold of 0.01 or 0.03 is popularly used to declare a SNV as being rare. Using microarrays, whole exome or whole genome DNA-sequencing methods (Goodwin et al., 2016; Mardis, 2017) allows to assess the variation of millions of SNVs, and genome-wide association studies (GWAS) have identified many common variants associated with a large number of complex traits, and explained some of the estimated heritability (Welter et al., 2014). The refinement of high-throughput technologies and the decrease in sequencing costs now allow investigating the role of rare variants in greater depth as well. However, despite evidence that many rare variants play a functional role in complex traits and in the regulation of biological processes (Gorlov et al., 2011; Nelson et al., 2012; Purcell et al., 2014), only a relatively small number of rare variants have been found to be associated with complex diseases so far (Hunt et al., 2013; Lohmueller et al., 2013; UK10K Consortium, 2015) with one likely reason that frequently used statistical tests have been underpowered.

For a more detailed description of the metabolic phenotype obesity, it can first be noted that obesity is defined through the amount of accumulated AT in the body. The most popular traditional anthropometric measures of obesity include the body mass index (BMI), waist and hip circumference (WC, HC), and the waist-to-hip ratio (WHR). However, they only provide surrogate measures of the body volume or body fat. Using a direct quantification of fat mass with computed tomography or MRI provides a more reliable and less biased quantification (Ross et al., 2003; Bosy-Westphal et al., 2008; Bredella et al., 2010; Taylor et al., 2012; Karlsson et al., 2013; Lee & Kuk, 2013) and also allows to differentiate between the different body fat compartments. In terms of volume, SAT and visceral AT (VAT) are the largest body fat compartments that constitute TAT. Regarding genetic association studies with cardiometabolic traits, there exist a number of recent large-scale meta-analyses with obesity traits such as BMI (Locke et al., 2015) or body fat distribution (BMI adjusted for waist-hip-ratio; Shungin et al., 2015), which have identified 97 and 49 associated genetic loci, respectively. Overall, there are currently 895 reported associations with BMI in the NHGRI-EBI GWAS Catalog (Hindorff et al., 2017; MacArthur et al., 2017), in addition to associations with other obesity traits.

For some background regarding the cardiovascular phenotype BP, the most popularly investigated traits include systolic and diastolic BP, mean arterial pressure, pulse pressure, and the hypertension status. While the BP traits are not directly tissue-based, tissues such as the kidney can be of specific interest in investigations into disease etiology (Tsai et al., 2017). Regarding the assessment of BP, there exist measure considerations (e.g., combining at least 3 subsequent measures to form a more robust mean BP measure) and statistical considerations, for example the adjustment of BP for antihypertensive medication intake requires special attention (Tobin et al., 2005; Konigorski et al., 2014). In the NHGRI-EBI GWAS Catalog, there are currently 263 reported genetic loci with associations to the above BP traits, which are driven to a large extent by recent large-scale GWAS and meta-analyses (Ehret et al., 2016; Liu et al., 2016; Surendran et al., 2016; Warren et al., 2017).

2.2 Existing statistical approaches for genetic association studies

In order to integrate and analyze complex genotype-phenotype data to answer specific biological questions, appropriate statistical methods with an efficient implementation are needed. One of the main methodological challenges is how to deal with the high dimensionality of the data ("curse of dimensionality"). The simplest statistical approach for any analysis associating genetic markers with phenotypes is to separately test each genetic marker for its association with each phenotype. However, the results have to be integrated later and the analyses cannot incorporate any dependencies between markers or traits. Furthermore, the multiple testing burden increases exponentially when multiple phenotypes are included. As a variation, genetic markers can be combined by using some form of aggregation or dimension reduction technique. After this dimension reduction step, the aggregated measures can be investigated in association tests, but it is unclear if this yields an improvement in general. Finally, statistical approaches can be used to analyze multiple traits simultaneously. This allows investigating more complex models and incorporating dependencies between variables and also increases the face validity of the results. More details are described in the following.

2.2.1 Models including multiple genetic markers

Aggregating genetic markers or traits under consideration can reduce the computational complexity and multiple testing burden. At the same time, however, such approaches rely on the assumption that the aggregated measures carry some biological meaning. Therefore, important considerations have to be whether the assumptions are biologically appropriate, and any comparison of MMTs versus SMTs has to consider that different hypotheses are tested. This will be illustrated for the context of genetic association studies in the following:

For GWAS, the unit of MMTs can be a region of the genome such as a gene or a pathway spanning multiple genes. Accordingly, MMTs are testing whether a given combination of the genetic markers in a given region ("burden-type tests") or any of the markers in a given region (variance-component-type tests) is associated with the trait of interest, whereas SMTs are testing whether a given marker is associated with the trait of interest. More formally, for a given region including k markers, the tested hypotheses for SMTs are

$$H_{0j} : \beta_j = 0 \quad \text{vs.} \quad H_{Aj} : \beta_j \neq 0 \quad \text{for } j = 1, \ldots, k \tag{2.2.1}$$

where β_j is the effect of the j-th marker on the trait under consideration; and the tested hypotheses for MMTs are

$$H_0 : \beta = 0 \quad \text{vs.} \quad H_A : \beta \neq 0 \tag{2.2.2}$$

where β is the effect of the region on the trait of interest, which could be a single parameter in burden-type tests, or a vector of parameters for the k marker effects in variance-component-type tests with

$$H_0 : \boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^T = 0 \quad \text{vs.} \quad H_A : \beta_j \neq 0 \text{ for at least one } j. \tag{2.2.3}$$

Hence, if SMTs and MMTs are compared by evaluating the power for identifying a given causal marker (SMT) and a given causal region (MMT), respectively, any conclusion that

one approach is more powerful than the other is questionable since different hypotheses are tested. For a valid comparison, their power has to be compared for identifying the same genetic locus, such as a gene.

Specific approaches how to aggregate multiple genetic markers have been a focus in the analysis of rare variants. Their development has been motivated by arguments that SMTs have inflated type I error and very low power when testing rare SNVs in the analysis of binary traits (Li & Leal, 2008), and might not provide valid statistical inference when testing very rare SNVs or single base-pair mutations. MMTs use different approaches to combine the rare genetic variants' information in a given region and test the association of the region with the phenotype. They can be broadly classified as (i) burden tests, which obtain genetic scores by collapsing rare variants in a region/gene (Morgenthaler & Thilly, 2007; Li & Leal, 2008; Madsen & Browning, 2009; Morris & Zeggini, 2010), (ii) extensions of burden tests (Price et al., 2010; Lin & Tang, 2011; Chen et al., 2012), (iii) variance-component tests, which collapse single-variant score statistics in a region (SKAT in Wu et al., 2011; and other tests in Basu & Pan, 2011, Neale et al., 2011), (iv) combinations of burden and variance-component tests (SKAT-O in Lee et al., 2012; aSPU in Pan et al., 2014, 2015; and other tests in Derkach et al., 2013; Sun et al., 2013), and (v) other approaches (Xu et al., 2012; Mieth et al., 2016). Detailed comparisons of rare variant association tests for binary traits (Madsen & Browning, 2009; Basu & Pan, 2011; Kinnamon et al., 2012; Xing et al., 2012; Pan et al., 2014) showed that SMTs (Score tests, Wald tests, and Fisher's exact tests) have deflated empirical type I errors and much lower power compared to MMTs across most considered scenarios. For the analysis of quantitative traits such as BP or obesity, however, extensive comparisons are lacking and few studies have compared SMTs with MMTs with inconclusive evidence (for more details, see Konigorski et al., 2017). Hence, a valid comparison of SMTs with MMTs for the analysis of quantitative traits, which is described in section 4.1, can have wide implications.

2.2.2 Models including genetic markers and multiple traits

In general, many different statistical methods have been developed for multivariate analyses, and can be grouped into supervised and unsupervised methods, parametric and nonparametric tests, and approaches from machine learning and from statistics. In this thesis, the focus is on genetic association analyses and hence on statistical methods that have been proposed therefore. Some of the proposed methods are based on SMTs and some on MMTs, and an overview about traditional and recently proposed methods for genetic association studies is given in the following in sections 2.2.2.1 (methods for joint models of multiple phenotypes) and 2.2.2.2 (methods to infer direct genetic effects incorporating multiple phenotypes). Many of these methods can be extended to the analysis of multi-level omics data (e.g., including genomic, transcriptomic, epigenomic, proteomic, metabolomic, and/or phenomic data), however, the analysis of such data has also brought forward specialized methods which have been summarized in, for example, Ritchie et al. (2015), Kristensen et al. (2014), and Gomez-Cabrero et al. (2014). Such methods often contain a larger focus on efficient implementation and exploratory analyses in contrast to a more detailed focus on specific statistical models and their interpretation such as in this thesis. Also, most of them focus on integrating mul-

tiple sets of predictors to predict one outcome rather than predicting multiple phenotypes, and are not further discussed here.

2.2.2.1 Genetic association analysis using joint models of multiple phenotypes

For genetic association analyses incorporating multiple phenotypes, most commonly, methods based on multivariate regression models, dimension reduction methods, or methods combining results from univariate analyses are employed. First, regarding regression-based approaches, multivariate analysis of variance (MANOVA), multivariate generalized linear mixed effects models (GLMMs), frailty models, or generalized estimating equation (GEE) models can be used to predict the effect of one or more genetic predictors on multiple phenotypes. In another approach, O'Reilly et al. (2012) proposed to invert the typical regression approach and to regress genetic markers on multiple phenotypes ("MultiPhen") for an SMT by using ordinal or multiple logistic regression approaches. Second, classical dimension reduction methods include principal component analysis and factor analysis (Pearson, 1901; Hotelling, 1933; Jolliffe, 2002; Bartholomew et al., 2011), structural equation modeling (SEM; Bollen, 1989) and canonical correlation analysis (CCA; Hotelling, 1936), which all allow extracting latent features from genotypes and from phenotypes and associating them. More specialized approaches targeting the explained variance by genetic variants have been developed in the form of principal component of heritability (PCH; Ott & Rabinovitz, 1999) and further extensions. An overview of methods from these two categories can also be found in Schillert & Konigorski (2016). Third, methods combining results from univariate analyses generally derive the distribution of a weighted linear combination of univariate test statistics to test the joint effect on multiple phenotypes. Proposed approaches include the O'Brien method (O'Brien, 1984) and methods described in Xu et al. (2003b), Zhang (2005) or Pan (2009). Instead of combining test statistics, TATES (van der Sluis et al., 2013) combines p-values from testing the association with multiple traits and derives a p-value for testing the joint association with all traits by correcting for the dependence between traits. In addition to this non-exhaustive list, many further approaches and extensions have been proposed. An overview about applications of machine learning approaches to learn associations between genetic markers and (multiple) phenotypes is given in Szymczak et al. (2009), and examples of Bayesian methods for association studies with multiple phenotypes can be found in Bottolo et al. (2013), Xu et al. (2014) and Jiang et al. (2015), but they are not further described here.

Overviews and reviews about the different approaches are given in Yang & Wang (2012) and Shriner (2016), and comparisons through simulation studies can be found in Galesloot et al. (2014), Zhang et al. (2014), Zhu et al. (2015) or Liang et al. (2016). The comparisons indicate that there is not one uniformly optimal test and that under different scenarios, different tests have the highest power. One major difference between the different approaches is whether they perform a multiple degree-of-freedom (df) test to evaluate the overall genetic association with all traits, or whether they perform single-df tests testing the genetic association with each trait separately. While multi-df tests generally have the highest power when a genetic variant affects all tested traits, they provide less or no information on which traits are relevant, and they loose power when only one or a few traits are affected. Further differences

include whether the approaches are able to model different dependence structures and trait distributions, and whether the dependences are rather treated as nuisance or as additional information.

In general, all of the above methods can be applied to association analyses of rare variants as well, however, the performance and validity of many approaches have not been described. Specialized rare-variant tests incorporating multiple phenotypes have been proposed in MURAT (Sun et al., 2016a, 2016b) and the multivariate aSPU and aSPUset tests (Zhang et al., 2014; Kim et al., 2016). MURAT is a multivariate extension of SKAT (Wu et al., 2011) yielding a gene-based MMT. aSPU and aSPUset are multivariate extensions of the univariate aSPU test described in Pan et al. (2014, 2015), which is an adaptive test combining different powered score tests. aSPU performs a multivariate SMT for each marker, aSPUset performs a MMT of all markers in a gene with a variation to additionally include a GEE Score test in the derived test statistic ("aSPUset-Score"). The different tests contained in the class of powered score tests have great similarity or encompass the CCA and MANOVA tests, which are in turn similar to MultiPhen, TATES, and other tests (Zhang et al., 2014). These tests are considered in the evaluation of C-JAMP in section 4.2. Other tests have been recently proposed in, for example, Broadaway et al. (2016), Liang et al. (2016), Wang et al. (2016) and Kaakinen et al. (2017), but are not further considered in this thesis.

2.2.2.2 Models to infer direct and indirect genetic effects

In addition to the methods described above, which mainly focus on overall genetic effects on multiple phenotypes, another class of approaches has been developed for the estimation and testing of direct genetic effects. For some background regarding their importance, as mentioned above, many genetic associations have been identified, for example, with obesity traits and type 2 diabetes (Fuchsberger et al., 2016; Locke et al., 2015). Some of these genetic markers are associated with multiple anthropometric traits (Ried et al., 2016), anthropometric and metabolic traits (Pickrell et al., 2016), and birthweight and type 2 diabetes (Zeng et al., 2017). However, it is unknown if these studies, and association studies in general, truly show evidence of functional genetic effects (e.g., through genetically-determined circulating biomarkers on type 2 diabetes, Lotta et al., 2016, or coronary artery disease, Helgadottir et al., 2016), of pleiotropic genetic effects on multiple phenotypes, or if the observed associations are due to indirect effects through some other intermediate phenotypes. Also, the genetic effects might be mediated or confounded by regulatory factors such as epigenetic markers (Feil & Fraga, 2012; Relton & Davey Smith, 2012a, 2012b; Corradin et al., 2016). As an example, Vansteelandt and colleagues (2009) showed that the effect estimate of a previously found association between a genetic marker and lung function was biased and could not be confirmed when the indirect effect of the genetic marker through weight was removed. For an example regarding BP, the association of genetic markers might be partially mediated by intermediate phenotypes such as, for example, gene expression (Huan et al., 2015a, 2015b). In addition to falsely identified markers, the direct genetic effects can also be masked in traditional statistical methods when there are indirect effects or confounded indirect effects in opposing direction of the direct effect. This background highlights the importance of using appropriate statistical methods that help disentangling direct and indirect genetic effects -

or more general molecular effects - through intermediate phenotypes, and if possible with robustness against effects of measured and unmeasured confounding factors.

For an overview of different statistical approaches, consider the situation of one genetic marker, one primary phenotype, one or more intermediate phenotypes and additional factors. Two traditional methods are (i) to include the genetic marker, intermediate phenotypes and factors as covariates in a multiple regression model of the primary phenotype, or (ii) to first regress the primary phenotype on the intermediate phenotypes and factors, and then regress the extracted residuals on the genetic marker (regression of residuals). The two approaches are frequently used for the analysis of primary phenotypes that are continuous (using a linear regression model; LM), and the multiple regression approach is frequently used for the analysis of binary or categorical primary phenotypes (using generalized linear regression models), or potentially censored time-to-event primary phenotypes (using, for example, proportional hazards (PH) or accelerated failure time (AFT) regression models). However, both traditional approaches can lead to biased point estimates and invalid testing of direct genetic effects on the primary phenotype in some situations, by removing part of the true association or by failing to remove the effect of the intermediate phenotype (i.e., of the indirect genetic effect) or unmeasured confounders (Rosenbaum, 1984; Cole & Hernán, 2002; Goetgeluk et al., 2008; Vansteelandt et al., 2009).

More elaborate approaches have been proposed to overcome these limitations, including the SEM method described in the previous section. Further approaches have been developed in studies on causal inference using structural nested models and G-estimation methods (Robins, 1986, 1992; Robins & Greenland, 1994; Goetgeluk et al., 2008), or the inverse probability weighting method (Robins et al., 2000). A more detailed overview of these different approaches can be found in Vansteelandt & Joffe (2014). Applications of the sequential G-estimation method to DAGs for estimating and testing direct genetic effects have been described for quantitative (Vansteelandt et al., 2009) and time-to-event primary phenotypes (using PH and AFT regression models, Lipman et al., 2011, and Aalen additive hazard models, Martinussen et al., 2011). These approaches include two steps: first, an adjusted phenotype is obtained by removing the effect of the intermediate phenotype from the primary phenotype. Then, the association of the genetic marker with the adjusted phenotype is tested by accounting for the additional variability obtained due to the estimation in the first stage in addition to the second stage (see section 3.2.1.1 for a formal description). Large-sample results of the estimator are provided for the analysis of Aalen additive hazard models in Martinussen et al. (2011). However, the focus of the sequential G-estimation methods for linear quantitative and log-linear time-to-event primary phenotype models was on testing the absence of direct genetic effects, and the standard errors and confidence intervals of the direct genetic effects were not investigated. Finally, the extension of the sequential G-estimation method described for time-to-event primary phenotypes using the PH and AFT regression models (Lipman et al., 2011) is invalid as will be shown in sections 4.3 and A.4.3.

2.3 Computational aspects

For measuring genetic data for the analysis, different technologies are available to provide in-depth information on a base-pair resolution, and sequencing technologies are currently a popular choice. In the bioinformatics processing steps, it is important that efficient and robust pipelines are used to extract, process and filter, quality-control and transform the relevant information from the raw data to the final datasets for the main analysis. For this, many different software tools and workflows are available for each processing step including the preprocessing, mapping to the reference genome, post-alignment processing and variant calling for whole-genome or whole-exome DNA-sequencing, and preprocessing, transcriptome assembly construction (if necessary) and read alignment, post-alignment processing and expression quantification for RNA-Seq analysis, followed by respective quality control and data analysis tools. Overviews and reviews are given in, for example, Bao et al. (2014), Pirooznia et al. (2014) and Hwang et al. (2015) for variant calling and DNA-sequencing, and Yang & Kim (2015), Conesa et al. (2016) and Everaert et al. (2017) for RNA-sequencing. Here, in the data analysis in section 5.2, the `Tophat-HTSeq` pipeline is used for read mapping and gene expression quantification, and `samtools` (Li et al., 2009; Li, 2011) is used for variant calling.

In the analysis of the generated high-dimensional genotype-phenotype data, whether a statistical method can be used depends on an efficient and robust implementation. Separate analyses of each genetic marker with each phenotype have a very low computational complexity and can be easily parallelized. In a joint modeling of a large number of genetic markers and phenotypes, one high-dimensional model has to be fitted. Here, efficient reformulations of the statistical model (e.g., for GLMMs, Lippert et al., 2011), approximations (Listgarten et al., 2012; Mandt et al., 2017) or a parallelization of the computation can help to scale the approach with the number of analyzed individuals (e.g., from cubic or square scaling to linear scaling; Lippert et al., 2011) and make it applicable for the analysis. As further help to reduce the computational and statistical testing burden, biological restrictions can be imposed on the analysis through prespecified assumptions regarding the relationships between phenotypes and SNVs. For example, the search for expression quantitative trait loci (eQTL) can be restricted to cis-acting loci (Konigorski et al., 2016).

For the two proposed methods in this thesis, parameter and standard error estimates for C-JAMP are obtained using the quasi-Newton, variable metric BFGS (Broyden-Fletcher-Goldfarb-Shanno; Gentle, 2009) method and from the inverse Fisher information matrix. For CIEE, the efficient least squares (LS) estimation and derived closed-form solutions are used. Since Wald tests are employed for hypothesis testing of the parameter estimates, there are no further computational costs. Both approaches are employed as single-marker tests so that they can be easily parallelized and in all analyses reported in this thesis, the separate analysis of all SNVs was spread over up to 500 cores at the Max Delbrück Center (MDC) computing cluster. This allowed to obtain all results of the data applications and simulation studies. Especially in the simulation studies, hundreds of millions of replicates had to be analyzed in order to obtain empirical type I error and power estimates under all scenarios, which led to a high computational burden with computations taking more than 300 computing years.

Chapter 3

Developed statistical methods

3.1 C-JAMP: Copula-based joint analysis of multiple phenotypes

The general goal of C-JAMP (copula-based modeling of multiple traits of a phenotype) is to jointly model two (or more) traits of a phenotype conditional on a genetic marker of interest using copula functions, in order to estimate and test the association of the marker with either trait. Here, the underlying argument is that through the joint modeling of multiple traits, the power of the association test with a given trait can be increased by using the information from the genetic association with the other traits and their dependence to increase the efficiency of the estimation.

For an intuition about the potential power increase when analyzing two traits jointly instead of one trait, consider the following simple scenario with two normally-distributed traits Y_1, Y_2 and a genetic marker X. Based on a simple linear regression model,

$$Y_1 = \alpha_0 + \alpha_{XY} x + \varepsilon,\ \varepsilon \sim N(0, \sigma^2),$$

the null hypothesis $H_0 : \alpha_{XY} = 0$ vs. $H_A : \alpha_{XY} \neq 0$ can be tested using the Wald-type test statistic $W = \frac{\widehat{\alpha}_{XY}}{\widehat{SE}(\widehat{\alpha}_{XY})}$ with $\widehat{SE}(\widehat{\alpha}_{XY}) = \sqrt{\frac{1}{n-2} \cdot \frac{\sum \widehat{\varepsilon}_i^2}{\sum (x_i - \bar{x})^2}}$. In a joint model of Y_1 and Y_2 given x, an upper bound to the power can be thought of as doubling the sample size, hence appending Y_1 with the second trait Y_2 and similarly appending the predictor values x with x. Then, the standard error estimate of $\widehat{\alpha}_{XY}$ would be $\widehat{SE}(\widehat{\alpha}_{XY}) = \sqrt{\frac{1}{2n-2} \cdot \frac{\sum \widehat{\varepsilon}_i^2}{\sum (x_i - \bar{x})^2}}$. Hence, in the joint model, the standard error could be maximally decreased by about $\sqrt{1/2}$ (i.e., $\approx 30\%$ smaller). In general, for a joint modeling of p traits, the upper bound on the standard error decrease is therefore of the order $\sqrt{1/p}$.

3.1.1 Statistical details

In the following, copula functions and C-JAMP will be described in more detail. Copulas are functions used to construct a joint distribution by combining the marginal distributions with a dependence structure (Joe, 1997; Nelsen, 2006). In contrast to multivariate normal models, which model the linear dependence of two random variables coming from a bivariate

normal distribution with a certain Pearson's correlation, more general dependencies can be flexibly investigated using copula functions. This can be helpful when the conditional mean of Y_1 given Y_2 is not linear in Y_2. In addition, transformations of Y_1 and Y_2 can change the value of Pearson's correlation but copulas based on a nonparametric correlation measure are invariant under strictly increasing transformations of the margins. Additional properties of copula models are that they allow the investigation of the dependence structure between the phenotypes separately from the marginal distributions. Also, the marginal distributions can come from different families and not necessarily from the normal distribution. Copula models can then be used to make inference about the marginal parameters and to identify genetic markers associated with one or multiple phenotypes, which will be the main aim explored here. They can also be used to make inference about the copula dependence parameters, for example, for an alternative definition of pleiotropy and to identify genetic variants that explain the dependence between two phenotypes (Konigorski et al., 2014). More formally:

Definition 3.1.1 (Copula)**.** A p-dimensional *copula* is defined as a function $C : [0,1]^p \rightarrow [0,1]$ which is a joint cumulative distribution function with uniform marginal cumulative distribution functions.

Hence, C is the distribution of a multivariate random vector, and can be considered independent of the margins. For $Y_1, \ldots, Y_p$ and a covariate vector $\boldsymbol{z}$, the joint distribution F of $Y_1, \ldots, Y_p$, conditional on $\boldsymbol{z}$, can be constructed by combining the marginal distributions of $Y_1, \ldots, Y_p$, $F_1, \ldots, F_p$, conditional on $\boldsymbol{z}$, using a copula function C_ψ with dependence parameter(s) ψ:

$$F(Y_1, \ldots, Y_p|\boldsymbol{z}) = C_\psi(F_1(Y_1|\boldsymbol{z}), \ldots, F_p(Y_p|\boldsymbol{z})). \tag{3.1.1}$$

For continuous marginal distributions, the copula C_ψ is unique and the multivariate distribution can be constructed from the margins in a uniquely defined way (Sklar, 1959):

Theorem 3.1.2 (Sklar, 1959)**.** *Let F be a p-dimensional distribution function with univariate margins $F_1, \ldots, F_p$. Then, there exists a copula C_ψ such that for all $\boldsymbol{z} = (z_1, \ldots, z_p)^T \in \mathbb{R}^p$, $F(z_1, \ldots, z_p) = C_\psi(F_1(z_1), \ldots, F_p(z_p))$. C_ψ is uniquely determined on the joint range of all F_l, i.e. on $F_1(\mathbb{R}) \times \cdots \times F_p(\mathbb{R})$. Hence, it is unique if $F_1, \ldots, F_p$ are continuous.*

The following theorem implies the converse, which is used in the construction of multivariate distributions:

Theorem 3.1.3. *Let $F_1, \ldots, F_p$ be univariate distribution functions and C_ψ a p-dimensional copula function. Then, $F : \mathbb{R}^p \rightarrow [0,1]$ defined through $F(z_1, \ldots, z_p) = C_\psi(F_1(z_1), \ldots, F_p(z_p))$ is a p-dimensional distribution function with margins $F_1, \ldots, F_p$.*

There exist several proofs of Theorem 3.1.2 using different approaches, for example in Schweizer & Sklar (1974) and Carley & Taylor (2002) for 2-dimensional copulas. Proofs for the general p-dimensional case, also covering the case when one or more margins are discrete, can be found in Burchard & Hajaiej (2006) and Rüschendorf (2009). While a detailed proof is not within the scope of this thesis, some intuition can be gained for the case of continuous margins, where C_ψ can be determined as $C_\psi = F\left(F_1^{[-1]}(z_1), \ldots, F_p^{[-1]}(z_p)\right)$ for all $\boldsymbol{z} \in (0,1)^p$ with $F_l^{[-1]}$ being the quasi-inverse of F_l, $F_l^{[-1]}(t) := \inf(\boldsymbol{z} : F_l(\boldsymbol{z}) \geq t)$.

Popular copula functions include the Clayton family (Clayton, 1978)

$$C_\varphi(u_1,\dots,u_p,\varphi) = \left(\left(\sum_{l=1}^{p} u_l^{-\varphi}\right) - (p-1)\right)^{-1/\varphi} \tag{3.1.2}$$

with $\varphi > 0$, and the Gumbel-Hougaard family (Gumbel, 1960)

$$C_\theta(u_1,\dots,u_p,\theta) = \exp\left(-\left[\sum_{l=1}^{p}(-\log(u_l))^\theta\right]^{1/\theta}\right) \tag{3.1.3}$$

with $\theta > 1$. A third family which includes both (3.1.2) and (3.1.3) for $\theta = 1$ and $\varphi \to 0$ is the 2-parameter copula family

$$C_\psi(u_1,\dots,u_p,\varphi,\theta) = \left(\left[\sum_{l=1}^{p}\left(u_l^{-\varphi}-1\right)^\theta\right]^{1/\theta} + 1\right)^{-1/\varphi} \tag{3.1.4}$$

with $0 \le u_l, u_{l'} \le 1$, and the copula parameters $\boldsymbol{\psi} = (\varphi,\theta)^T, \varphi > 0, \theta \ge 1$, which allows a flexible modeling of both the lower- and upper-tail dependence. All three above families are examples of Archimedean copulas (Genest & Rivest, 1993), which are defined as follows.

Definition 3.1.4 (Archimedean copula)**.** Let ξ_ψ be a decreasing convex function on the unit interval with $\xi_\psi(1) = 0$. A copula C_ψ is called *Archimedean*, if it can be represented as

$$C_\psi(u_1,\dots,u_p,\psi) = \xi_\psi^{-1}(\xi_\psi(u_1) + \dots + \xi_\psi(u_p)). \tag{3.1.5}$$

For an investigation of the dependence between the traits, different dependence measures can be derived from the copula parameters. In the description and evaluation of C-JAMP, Kendall's τ is used as measure of dependence, which is the probability of concordance minus the probability of discordance. For the copula family in (3.1.4), it can be derived as (Joe, 1997)

$$\tau = 1 - \frac{2}{\theta(\varphi+2)}. \tag{3.1.6}$$

A visual illustration of the different dependence structures captured by the Clayton copula is shown in Figure 3.1.

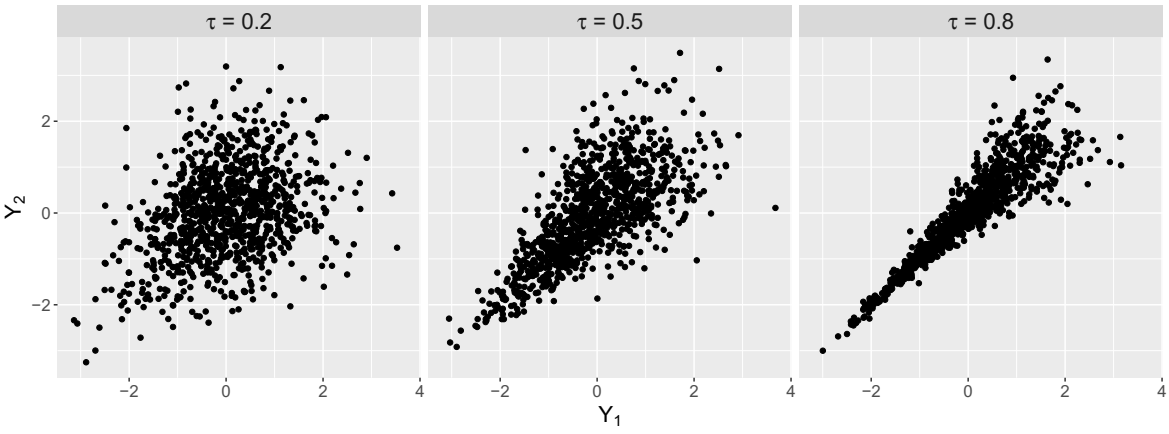


Figure 3.1: Scatterplots of bivariate data (Y_1 and Y_2) from Clayton's copula with dependence Kendall's $\tau = 0.2, 0.5, 0.8$ and standard normal margins.

Additional dependence parameters of interest can be the lower and upper tail dependence (λ_L, λ_U, respectively), which explain the amount of dependence between extreme values, and can give more insight in identifying pleiotropic variants. For the copula family in (3.1.4), these dependence measures are (Joe, 1997)

$$\lambda_L = 2^{-\frac{1}{\theta\varphi}}, \quad \lambda_U = 2 - 2^{\frac{1}{\theta}}.$$

In the following, C-JAMP is presented for the analysis of two quantitative, normally-distributed traits Y_1 and Y_2, which are modeled conditional on a covariate vector $\boldsymbol{z} = (z_1, z_2, x)^T$ including one genetic marker X and covariates Z_1 and Z_2. As marginals, linear models are considered:

$$Y_1 = \alpha_0 + \alpha_1 z_1 + \alpha_2 z_2 + \alpha_{XY} x + \varepsilon, \tag{3.1.7}$$

$$Y_2 = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_{XY} x + \varepsilon', \tag{3.1.8}$$

with the parameter vectors $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2, \alpha_{XY})^T$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_{XY})^T$ and $\varepsilon \sim N(0, \sigma_1^2)$, $\varepsilon' \sim N(0, \sigma_2^2)$, under the assumption that the covariates are mutually independent. The fact that the parameters measuring the dependence between Y_1 and Y_2 do not appear in the marginal distributions is very useful for estimating the effect of a genetic marker X on Y_1 or Y_2 while considering the dependence between Y_1 and Y_2.

Focusing on fully parametric models, maximum likelihood estimates (MLEs) $\widehat{\boldsymbol{\omega}}$ of the parameter vector $\boldsymbol{\omega} = (\boldsymbol{\psi}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma_1, \sigma_2)^T$ can be obtained by maximizing the likelihood function of the data in (3.1.9), which is derived in Lemma 3.1.5 for the 2-parameter copula and used in the implementation of C-JAMP. Alternatively, two-stage pseudo-likelihood methods could be used (Shih & Louis, 1995; Joe, 1997).

$$\begin{aligned} L(\boldsymbol{\omega}) &= \prod_{i=1}^{n} \left[\frac{\partial^2 F(y_{1i}, y_{2i} | z_{1i}, z_{2i}, x_i)}{\partial y_{1i} \partial y_{2i}} \right] = \prod_{i=1}^{n} \left[\frac{\partial^2 C_\psi \left(F_1(y_{1i} | z_{1i}, z_{2i}, x_i), F_2(y_{2i} | z_{1i}, z_{2i}, x_i) \right)}{\partial y_{1i} \partial y_{2i}} \right] \\ &= \prod_{i=1}^{n} \left[\frac{\partial^2 C_\psi \left(F_1(y_{1i} | z_{1i}, z_{2i}, x_i), F_2(y_{2i} | z_{1i}, z_{2i}, x_i) \right)}{\partial F_1(y_{1i} | z_{1i}, z_{2i}, x_i) \, \partial F_2(y_{2i} | z_{1i}, z_{2i}, x_i)} \right] \\ &\quad \times \left[\frac{\partial F_1(y_{1i} | z_{1i}, z_{2i}, x_i)}{\partial y_{1i}} \right] \times \left[\frac{\partial F_2(y_{2i} | z_{1i}, z_{2i}, x_i)}{\partial y_{2i}} \right]. \end{aligned} \tag{3.1.9}$$

Lemma 3.1.5. *Consider the 2-parameter copula function in (3.1.4) with marginal models in (3.1.7), (3.1.8). Then, the likelihood can be derived as*

$$\begin{aligned} L(\boldsymbol{\omega}) = \prod_{i=1}^{n} & f_1 f_2 {F_1}^{-\varphi-1} {F_2}^{-\varphi-1} \left[{F_1}^{-\varphi} - 1 \right]^{\theta-1} \left[{F_2}^{-\varphi} - 1 \right]^{\theta-1} \left[\left(\left({F_1}^{-\varphi} - 1 \right)^{\theta} + \left({F_2}^{-\varphi} - 1 \right)^{\theta} \right)^{\frac{1}{\theta}} + 1 \right]^{-\frac{1}{\varphi}-2} \\ & \cdot \left[\left({F_1}^{-\varphi} - 1 \right)^{\theta} + \left({F_2}^{-\varphi} - 1 \right)^{\theta} \right]^{\frac{1}{\theta}-2} \cdot \left((1 + \varphi\theta) \left[\left({F_1}^{-\varphi} - 1 \right)^{\theta} + \left({F_2}^{-\varphi} - 1 \right)^{\theta} \right]^{\frac{1}{\theta}} + (\theta - 1)\varphi \right) \end{aligned}$$

with $f_1 = \frac{1}{\sigma_1} \phi \left(\frac{y_{1i} - \alpha_0 - \alpha_1 z_{1i} - \alpha_2 z_{2i} - \alpha_{XY} x_i}{\sigma_1} \right)$, $f_2 = \frac{1}{\sigma_2} \phi \left(\frac{y_{2i} - \beta_0 - \beta_1 z_{1i} - \beta_2 z_{2i} - \beta_{XY} x_i}{\sigma_2} \right)$, *as well as* $F_1 = \Phi \left(\frac{y_{1i} - \alpha_0 - \alpha_1 z_{1i} - \alpha_2 z_{2i} - \alpha_{XY} x_i}{\sigma_1} \right)$ *and* $F_2 = \Phi \left(\frac{y_{2i} - \beta_0 - \beta_1 z_{1i} - \beta_2 z_{2i} - \beta_{XY} x_i}{\sigma_2} \right)$, *where* ϕ *is the standard normal probability density function (PDF) and* Φ *is the standard normal cumulative distribution function (CDF).*

Proof Lemma 3.1.5:

$$
\begin{aligned}
L(\boldsymbol{\omega}) &= \prod_{\mathrm{i}=1}^{\mathrm{n}} \frac{\partial^2 \left\{ \left[\left(F_1{}^{-\varphi}-1\right)^{\theta} + \left(F_2{}^{-\varphi}-1\right)^{\theta} \right]^{\frac{1}{\theta}} + 1 \right\}^{-\frac{1}{\varphi}}}{\partial F_1 \partial F_2} \cdot f_1 \cdot f_2 \\
&= \prod_{\mathrm{i}=1}^{\mathrm{n}} \frac{\partial \left\{ \left[\left(\left(F_1{}^{-\varphi}-1\right)^{\theta} + \left(F_2{}^{-\varphi}-1\right)^{\theta} \right)^{\frac{1}{\theta}} + 1 \right]^{-\frac{1}{\varphi}-1} \left[\left(F_1{}^{-\varphi}-1\right)^{\theta} + \left(F_2{}^{-\varphi}-1\right)^{\theta} \right]^{\frac{1}{\theta}-1} \left[F_1{}^{-\varphi}-1\right]^{\theta-1} \left[F_1{}^{-\varphi-1}\right] \right\}}{\partial F_2} \cdot f_1 \cdot f_2 \\
&= \prod_{i=1}^{n} f_1 f_2 \left[F_1{}^{-\varphi}-1\right]^{\theta-1} F_1{}^{-\varphi-1} \left(\left\{ \frac{\partial \left\{ \left[\left(\left(F_1{}^{-\varphi}-1\right)^{\theta} + \left(F_2{}^{-\varphi}-1\right)^{\theta} \right)^{\frac{1}{\theta}} + 1 \right]^{-\frac{1}{\varphi}-1} \right\}}{\partial F_2} \left[\left(F_1{}^{-\varphi}-1\right)^{\theta} + \left(F_2{}^{-\varphi}-1\right)^{\theta} \right]^{\frac{1}{\theta}-1} \right\} \right. \\
&\qquad \left. + \left\{ \left[\left(\left(F_1{}^{-\varphi}-1\right)^{\theta} + \left(F_2{}^{-\varphi}-1\right)^{\theta} \right)^{\frac{1}{\theta}} + 1 \right]^{-\frac{1}{\varphi}-1} \frac{\partial \left[\left(F_1{}^{-\varphi}-1\right)^{\theta} + \left(F_2{}^{-\varphi}-1\right)^{\theta} \right]^{\frac{1}{\theta}-1}}{\partial F_2} \right\} \right) \\
&= \prod_{i=1}^{n} f_1 f_2 \left[F_1{}^{-\varphi}-1\right]^{\theta-1} F_1{}^{-\varphi-1} \left(\left\{ \left\{ (1+\varphi) \left[\left(\left(F_1{}^{-\varphi}-1\right)^{\theta} + \left(F_2{}^{-\varphi}-1\right)^{\theta} \right)^{\frac{1}{\theta}} + 1 \right]^{-\frac{1}{\varphi}-2} \left[\left(F_1{}^{-\varphi}-1\right)^{\theta} + \left(F_2{}^{-\varphi}-1\right)^{\theta} \right]^{\frac{1}{\theta}-1} \right. \right. \right. \\
&\qquad \left. \left. \cdot \left[F_2{}^{-\varphi}-1\right]^{\theta-1} \left[F_2{}^{-\varphi-1}\right] \right\} \cdot \left[\left(F_1{}^{-\varphi}-1\right)^{\theta} + \left(F_2{}^{-\varphi}-1\right)^{\theta} \right]^{\frac{1}{\theta}-1} \right\} + \left\{ \left[\left(\left(F_1{}^{-\varphi}-1\right)^{\theta} + \left(F_2{}^{-\varphi}-1\right)^{\theta} \right)^{\frac{1}{\theta}} + 1 \right]^{-\frac{1}{\varphi}-1} \right. \\
&\qquad \left. \left. \cdot \left\{ (\theta-1)\,\varphi \cdot \left[\left(F_1{}^{-\varphi}-1\right)^{\theta} + \left(F_2{}^{-\varphi}-1\right)^{\theta} \right]^{\frac{1}{\theta}-2} \left[F_2{}^{-\varphi}-1\right]^{\theta-1} \left[F_2{}^{-\varphi-1}\right] \right\} \right\} \right) \\
&= \prod_{i=1}^{n} f_1 f_2 F_1{}^{-\varphi-1} F_2{}^{-\varphi-1} \left[F_1{}^{-\varphi}-1\right]^{\theta-1} \left[F_2{}^{-\varphi}-1\right]^{\theta-1} \left[\left(\left(F_1{}^{-\varphi}-1\right)^{\theta} + \left(F_2{}^{-\varphi}-1\right)^{\theta} \right)^{\frac{1}{\theta}} + 1 \right]^{-\frac{1}{\varphi}-2} \\
&\qquad \cdot \left[\left(F_1{}^{-\varphi}-1\right)^{\theta} + \left(F_2{}^{-\varphi}-1\right)^{\theta} \right]^{\frac{1}{\theta}-2} \cdot \left(\left\{ (1+\varphi) \left[\left(F_1{}^{-\varphi}-1\right)^{\theta} + \left(F_2{}^{-\varphi}-1\right)^{\theta} \right]^{\frac{1}{\theta}} \right\} \right. \\
&\qquad \left. + \left\{ (\theta-1)\,\varphi \left[\left(\left(F_1{}^{-\varphi}-1\right)^{\theta} + \left(F_2{}^{-\varphi}-1\right)^{\theta} \right)^{\frac{1}{\theta}} + 1 \right] \right\} \right) \\
&= \prod_{i=1}^{n} f_1 f_2 F_1{}^{-\varphi-1} F_2{}^{-\varphi-1} \left[F_1{}^{-\varphi}-1\right]^{\theta-1} \left[F_2{}^{-\varphi}-1\right]^{\theta-1} \left[\left(\left(F_1{}^{-\varphi}-1\right)^{\theta} + \left(F_2{}^{-\varphi}-1\right)^{\theta} \right)^{\frac{1}{\theta}} + 1 \right]^{-\frac{1}{\varphi}-2} \\
&\qquad \cdot \left[\left(F_1{}^{-\varphi}-1\right)^{\theta} + \left(F_2{}^{-\varphi}-1\right)^{\theta} \right]^{\frac{1}{\theta}-2} \cdot \left((1+\varphi\theta) \left[\left(F_1{}^{-\varphi}-1\right)^{\theta} + \left(F_2{}^{-\varphi}-1\right)^{\theta} \right]^{\frac{1}{\theta}} + (\theta-1)\,\varphi \right).
\end{aligned}
$$

□

Under some regularity conditions and the assumption that the model is correct, the MLEs $\widehat{\boldsymbol{\omega}}$ are consistent estimates of $\boldsymbol{\omega}$ and $\sqrt{n}\,(\widehat{\boldsymbol{\omega}} - \boldsymbol{\omega}) \xrightarrow{d} \mathrm{MVN}(\mathbf{0}, \boldsymbol{I}^{-1}(\boldsymbol{\omega}))$ in distribution, where $\boldsymbol{I}(\boldsymbol{\omega})$ is the Fisher information matrix $\boldsymbol{I}(\boldsymbol{\omega}) = E\left[-\frac{\partial^2 \mathrm{logL}(\boldsymbol{\omega})}{\partial \boldsymbol{\omega} \partial \boldsymbol{\omega}^T}\right]$. Therefore, variance estimates for the MLEs $\widehat{\boldsymbol{\omega}} = (\widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\sigma}_1, \widehat{\sigma}_2)$ in (3.1.9) can be obtained from the inverse of the observed information matrix using the Delta method, if necessary.

More specifically, in the implementation in R described in section 3.1.2, C-JAMP is described for the Clayton and 2-parameter copula models with normal marginal models, and parameter estimates are derived for $\log(\varphi), \log(\theta-1), \log(\sigma_1), \log(\sigma_2)$ and $\boldsymbol{\alpha}, \boldsymbol{\beta}$. Hence, the variance estimates from the Fisher information matrix are for these transformed parameters

and can be retransformed using the Delta method. For example, for the 2-parameter copula in (3.1.4) and the normal marginal models in (3.1.7)-(3.1.8) including the predictors Z_1, Z_2, X, the standard error estimates of the MLEs $\widehat{\boldsymbol{\omega}} = (\widehat{\varphi}, \widehat{\theta}, \widehat{\alpha}_0, \widehat{\alpha}_1, \widehat{\alpha}_2, \widehat{\alpha}_{XY}, \widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_{XY}, \widehat{\sigma}_1, \widehat{\sigma}_2)$ are:

$$
\begin{aligned}
\widehat{SE}(\widehat{\varphi}) &= \sqrt{[\boldsymbol{I^{-1}}(\boldsymbol{\omega})]_{1,1} \cdot \widehat{\varphi}^2}, \\
\widehat{SE}(\widehat{\theta}) &= \sqrt{[\boldsymbol{I^{-1}}(\boldsymbol{\omega})]_{2,2} \cdot (\widehat{\theta} - 1)^2}, \\
\widehat{SE}(\widehat{\alpha}_0) &= \sqrt{[\boldsymbol{I^{-1}}(\boldsymbol{\omega})]_{3,3}}, & \widehat{SE}(\widehat{\alpha}_1) &= \sqrt{[\boldsymbol{I^{-1}}(\boldsymbol{\omega})]_{4,4}}, \\
\widehat{SE}(\widehat{\alpha}_2) &= \sqrt{[\boldsymbol{I^{-1}}(\boldsymbol{\omega})]_{5,5}}, & \widehat{SE}(\widehat{\alpha}_{XY}) &= \sqrt{[\boldsymbol{I^{-1}}(\boldsymbol{\omega})]_{6,6}}, \\
\widehat{SE}(\widehat{\beta}_0) &= \sqrt{[\boldsymbol{I^{-1}}(\boldsymbol{\omega})]_{7,7}}, & \widehat{SE}(\widehat{\beta}_1) &= \sqrt{[\boldsymbol{I^{-1}}(\boldsymbol{\omega})]_{8,8}}, \\
\widehat{SE}(\widehat{\beta}_2) &= \sqrt{[\boldsymbol{I^{-1}}(\boldsymbol{\omega})]_{9,9}}, & \widehat{SE}(\widehat{\beta}_{XY}) &= \sqrt{[\boldsymbol{I^{-1}}(\boldsymbol{\omega})]_{10,10}}, \\
\widehat{SE}(\widehat{\sigma}_1) &= \sqrt{[\boldsymbol{I^{-1}}(\boldsymbol{\omega})]_{11,11} \cdot \widehat{\sigma}_1^2}, & \widehat{SE}(\widehat{\sigma}_2) &= \sqrt{[\boldsymbol{I^{-1}}(\boldsymbol{\omega})]_{12,12} \cdot \widehat{\sigma}_2^2},
\end{aligned}
$$

where $[\boldsymbol{I^{-1}}(\boldsymbol{\omega})]_{i,j}$ indicates the (i,j)-th element of the inverse Fisher information matrix.

To test whether the genetic marker X is associated with Y_1 and Y_2, large-sample methods based on the likelihood function can be used. In this thesis, it is proposed to test the null hypotheses $H_0 : \alpha_{XY} = 0$ (vs. $H_A : \alpha_{XY} \neq 0$) and $H_0 : \beta_{XY} = 0$ (vs. $H_A : \beta_{XY} \neq 0$), respectively, by using the large-sample Wald-type test statistics $\frac{\widehat{\alpha}_{XY}}{\widehat{SE}(\widehat{\alpha}_{XY})}$ and $\frac{\widehat{\beta}_{XY}}{\widehat{SE}(\widehat{\beta}_{XY})}$.

Finally, for testing goodness-of-fit of the copulas, the embedding the proposed copulas in an expanded family of copulas can be used. For example, to test the fit of copula models (3.1.2) and (3.1.3) (i.e., $H_0 : \theta = 1$ and $H_0 : \varphi \to 0$), the likelihood ratio test (LRT) statistic

$$
\text{logLR} = 2 \cdot \Big(L(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\psi}}) - L(\widehat{\boldsymbol{\alpha}}(\boldsymbol{\psi_0}), \widehat{\boldsymbol{\beta}}(\boldsymbol{\psi}_0), \widehat{\boldsymbol{\psi}}_0) \Big) \tag{3.1.10}
$$

can be used, where $\widehat{\boldsymbol{\psi}} = (\widehat{\varphi}, \widehat{\theta})$ and $\widehat{\boldsymbol{\psi}}_0 = (\widehat{\varphi}(\theta = 1), 1)$ when testing the Clayton copula, or $\widehat{\boldsymbol{\psi}}_0 = (0, \widehat{\theta}(\varphi = 0))$ when testing the G-H copula. Here, $\widehat{\boldsymbol{\alpha}}(\boldsymbol{\psi}_0)$ and $\widehat{\boldsymbol{\beta}}(\boldsymbol{\psi}_0)$ are the maximum likelihood estimates (MLEs) of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ under the null model $\boldsymbol{\psi} = \boldsymbol{\psi}_0$. Evaluating the likelihood ratio statistic (3.1.10) means testing parameters on the boundary of the parameter space where the usual chi-square approximation is inaccurate. However, it has been shown that (3.1.10) is asymptotically distributed as $0.5 + 0.5\chi^2_{(1)}$ under the null hypothesis $\boldsymbol{\psi} = \boldsymbol{\psi}_0$ and under the assumption that other free parameter in $\boldsymbol{\psi}_0$ don't lie on the boundary of the parameter space (Self & Liang, 1987).

3.1.2 Computational details & implementation

C-JAMP is implemented in the R package `CJAMP` for the joint analysis of two phenotypes Y_1 and Y_2 conditional of one or more predictors z with linear marginal models in (3.1.7), (3.1.8), for the Clayton and 2-parameter copula. `CJAMP` is currently available upon request as a bundled package. It can be installed through the zipped tar file with

```
install.packages("pathtofile/CJAMP_0.1.0.tar.gz", repos = NULL)
```

and then loaded with

```
library(CJAMP)
```

CJAMP was generated using RStudio Version 1.0.143 based on R version 3.4.1 (R Core Team, 2017) using the `devtools` package (Wickham & Chang, 2017). The documentation of the functions and generation of help pages was done using the `roxygen2` package (Wickham et al., 2017). Finally, `knitr` (Xie, 2017) and `rmarkdown` (Allaire et al., 2017) were used for the detailed long-form documentation (vignette), which is available through

```
browseVignettes("CJAMP")
```

More details for each function can be obtained through the regular help files, for example,

```
?cjamp
```

In the following, an outline of the implementation is given, as well as details regarding the application to genome-wide analyses. Examples for the use of the R functions are shown in the appendix in section A.7.1.

3.1.2.1 Data generation functions

Data Y_1, Y_2 can be sampled from the Clayton copula (with standard normal margins) using the `generate_clayton_copula()` function. This uses the steps described in Box 3.1 to generate Y_1, Y_2 from the Clayton copula $C_\varphi(\Phi(Y_1), \Phi(Y_2)) = C_\varphi(U_1, U_2)$ with standard normal marginal distributions, where Φ is the standard normal cumulative distribution function:

1. Generate uniform random variables $U_1, \tilde{U}_2 \sim \text{Unif}(0, 1)$.
2. Generate Y_1 from the inverse standard normal distribution of U_1, $Y_1 := \Phi^{-1}(U_1)$.
3. In order to obtain the second variable Y_2, use the conditional distribution of U_2 given U_1, set $\tilde{U}_2 = \frac{\partial C_\varphi(U_1, U_2)}{\partial U_1}$ and solve for U_2. This yields
$$U_2 = \left(1 - U_1^{-\phi} \cdot \left(1 - \tilde{U}_2^{-\phi/(1+\phi)}\right)\right)^{-1/\phi}.$$
4. Generate Y_2 from the inverse standard normal distribution of U_2, $Y_2 := \Phi^{-1}(U_2)$.

Box 3.1: Steps to generate data from the Clayton copula.

In order to generate sample data including SNVs and covariates, different functions are provided. The functions `generate_singleton_data()`, `generate_doubleton_data()` and `generate_genodata()` can be used to generate genetic data in the form of single nucleotide

variants. `generate_singleton_data()` generates singletons (i.e., SNVs with one observed minor allele); `generate_doubleton_data()` generates doubletons (i.e., SNVs with two observed minor alleles), and `generate_genodata()` generates n observations of k SNVs with random minor allele frequencies. The minor allele frequencies of the generated SNVs can be calculated using the function `compute_MAF()`.

Next, the functions `generate_phenodata_1_simple()`, `generate_phenodata_1()`, as well as `generate_phenodata_2_bvn()` and `generate_phenodata_2_copula()` can be used to generate phenotype data based on input SNVs, covariates, and a specification of the covariate effect sizes. Here, `generate_phenodata_1_simple()` and `generate_phenodata_1()` generate one normally-distributed or binary phenotype Y conditional on input SNVs and covariates X_1, X_2. `generate_phenodata_2_bvn()` and `generate_phenodata_2_copula()` generate two phenotypes Y_1, Y_2 with dependence Kendall's τ conditional on the SNVs and covariates X_1, X_2 from the bivariate normal distribution or the Clayton copula function with standard normal marginal distributions. `generate_phenodata_2_copula()` is using the function `generate_clayton_copula()` described above.

For genetic association analyses, the amount of phenotypic variance that can be explained by the SNVs might be of interest. This can be computed based on four different approaches using the function `compute_expl_var()` (see the help pages of the function for more details).

3.1.2.2 C-JAMP functions

C-JAMP is implemented for the joint analysis of two phenotypes Y_1, Y_2 conditional on one or more predictors $\boldsymbol{z}$. Functions are available to fit the joint model

$$F(Y_1, Y_2|\boldsymbol{z}) = C_{\boldsymbol{\psi}}(F_1(Y_1|\boldsymbol{z}),\ F_2(Y_2|\boldsymbol{z}))$$

for the Clayton and 2-parameter copula $C_{\boldsymbol{\psi}}$, with marginal models

$$Y_1 = \boldsymbol{\alpha}^T\boldsymbol{z} + \varepsilon_1,\ \varepsilon_1 \sim N(0, \sigma_1^2)$$
$$Y_2 = \boldsymbol{\beta}^T\boldsymbol{z} + \varepsilon_2,\ \varepsilon_2 \sim N(0, \sigma_2^2).$$

Naïve coefficient estimates of the parameter vector $\boldsymbol{\omega} = (\boldsymbol{\psi}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma_1, \sigma_2)^T$ can be obtained using the `get_estimates_naive()` function, which fits separate linear regression models of Y_1 and Y_2. For these or any other parameter estimates, the log-likelihood (or rather the minus log-likelihood) of the copula model can be computed with the function `minusloglik()` for the Clayton and 2-parameter copula. `minusloglik()` (and the `cjamp()`, `cjamp_loop()` functions below) use an additive model, which means that if categorical predictors are to be analyzed with more than 2 levels, then dummy variables have to be created beforehand. Accordingly, if single nucleotide variants (SNVs) are included as predictors, the computation is based on an additive genetic model if SNVs are provided as 0-1-2 genotypes and on a dominant model if SNVs are provided as 0-1 genotypes.

The `lrt_copula()` can be used to test the model fit of different nested copula models with the same marginal models using likelihood ratio tests. To test the fit of the 1-parameter Clayton copula model fit versus the 2-parameter copula (i.e., $H_0 : \phi \to 0$), the likelihood ratio test statistic in (3.1.10) is used with $\widehat{\boldsymbol{\psi}}_0 = (\widehat{\varphi}(\theta = 1), 1)$. P-values are obtained by using that $logLR$ is asymptotically distributed as $0.5 + 0.5\chi_1^2$ under the null hypothesis $\boldsymbol{\psi} = \boldsymbol{\psi}_0$

and under the assumption that other free parameters in $\boldsymbol{\psi}_0$ don't lie on the boundary of the parameter space (Self & Liang, 1987). Likelihood ratio tests of marginal parameters within the same copula model can be performed using the `lrt_param()` function.

Finally, the function `cjamp()` can be used to obtain MLEs of $\boldsymbol{\omega}$, standard error estimates, and perform hypothesis tests of the parameters, for the Clayton or 2-parameter copula. For maximizing the nonlinear log-likelihood function (i.e., minimizing `minusloglik()`), `cjamp()` uses the `optimx()` function of the `optimx` package. The parameters `SE_est` and `pval_est` allow to control whether standard error estimates are obtained from the observed inverse information matrix extracted from the optimization, and whether large-sample Wald-type hypothesis tests are performed to test the absence of the marginal genetic effects on both traits (i.e., $H_0 : \alpha_{XY} = 0$ and $H_0 : \beta_{XY} = 0$), as well as of the other marginal effects on each phenotype. Multiple predictors can be supplied, and if the goal is to test a large number of predictors sequentially with the same marginal models (such as in genetic association studies), the function `cjamp_loop()` provides an easy wrapper of the `cjamp()` function and only extracts the coefficient and standard error estimates as well as the p-values for the predictors of interest, and not for all other covariates. Both `cjamp()` and `cjamp_loop()` return `cjamp` objects as output, so that the written `summary.cjamp()` function can be used through the generic `summary()` to provide a reader-friendly output of the results.

For more technical details, preliminary investigations showed that the fastest and most robust convergence was achieved using the BFGS optimization method among the different available optimization algorithms, which is hence recommended. Different starting values on a grid are automatically tried in `cjamp()`, and the maximum number of tries can be controlled through the parameter `n_iter_max`. Some technical steps involve the multiplication of the second trait with -1 if the dependence between Y_1 and Y_2 is negative (since both the Clayton and 2-parameter copula can only fit positive dependencies), noted with a warning in the output for the interpretation. Furthermore, the parameter `scale_var` can be used to automatically scale all predictors. This is recommended, since the BFGS optimization can be sensitive to a different scaling of the predictors, which can result in a non-convergence of the optimization. In order to deal with convergence problems, several checks are built into `cjamp()`, such as checking the finiteness of the likelihood function as well as whether the `optim` convergence code indicates convergence and the Karush-Kuhn-Tucker (KKT) conditions are both satisfied. If any condition is not satisfied, the optimization is restarted with the next starting value until convergence or until the maximum number of `n_iter_max` is reached. In the latter case, standard errors and p-values are not computed.

3.1.2.3 C-JAMP application details

The multiple robustness checks and optimization attempts using different starting values increase the running time of `cjamp()`, but are necessary if the function is to be applied for a genome-wide association analysis - since in this case, manual checks are not possible. For GWAS, C-JAMP can be applied by either sequentially computing `cjamp()` on all SNVs of interest, or by supplying a matrix of multiple SNVs to the `cjamp_loop()` function. The running time of `cjamp()` increases with the number of covariates and the sample size, which is visualized in Figure 3.2. As can be expected, computing C-JAMP based on the 2-parameter

copula takes longer than based on the Clayton copula, since one more (dependence) parameter has to be estimated. As a concrete example, for a sample size of $n = 1000$, two traits Y_1 and Y_2, 1 SNV and 5 covariates, C-JAMP takes on average about 1.5 minutes running time to finish and obtain estimates as well as p-values. This is obviously much slower compared to univariate regression models, but still sufficient to allow for genome-wide analyses. A GWAS including 1,000,000 SNVs with a sample size of $n = 1000$, two traits Y_1 and Y_2, and 5 covariates can be computed in about 2 days on a cluster using 500 computing cores.

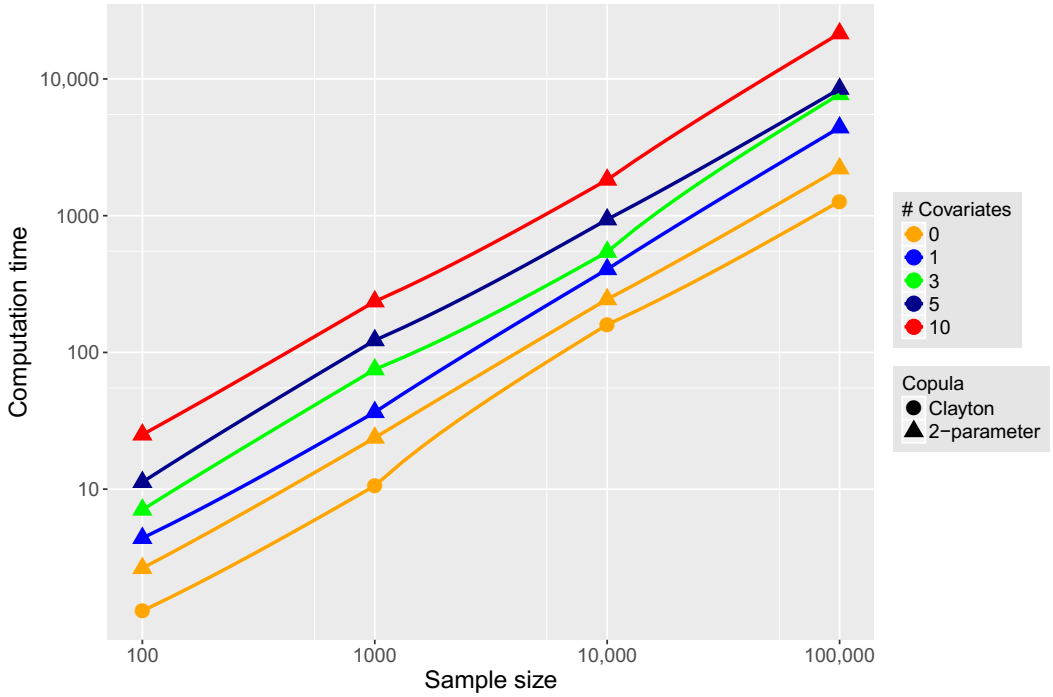


Figure 3.2: Computation time of C-JAMP. Shown are the mean computation times in seconds ('user times' obtained through the R `system.time()` function) of the `cjamp()` R function based on 10 runs, for a SNV with random MAF, using the Clayton or 2-parameter copula with marginal models as in equations (3.1.7)-(3.1.8) with different numbers of covariates, and data generated under model (4.2.1) with the effect size $c_{Y_1} = c_{Y_2} = 0.3$ of the causal SNV.

A more detailed illustration of the memory and running time used by the `cjamp()` function is shown in Figure 3.3, and highlights again that the computationally expensive part lies in the optimization.

Figure 3.3: Detailed computation time and memory use of C-JAMP. Shown are the memory use (in MB) and computation time (in ms) of each of the steps of the `cjamp()` R function, computed using the `profvis()` function of the `profvis` R package (Chang & Luraschi, 2017), using the Clayton or 2-parameter copula with marginal models as in equations (3.1.7)-(3.1.8), and data generated under model 4.2.1 with $n = 1000$ and the effect size $c_{Y_1} = c_{Y_2} = 0.3$ of the causal SNV.

3.2 CIEE: Causal inference using estimating equations

After the investigation of joint models using copula functions in the previous section, the focus in this section is on the estimation and testing of direct causal effects. For example, a given SNV might affect the gene expression level of gene s (eQTL) which further affects obesity traits, but the genetic effect on gene expression is dependent on other factors. Also, post-transcriptional regulation of the gene expression levels might play a role in their involvement in obesity. In such situations, while the indirect mediated effects are biologically interesting, a priori knowledge about the regulation and mediation would be necessary to model them. If the intermediate and confounding factors are not measured, the genetic effect can be obscured. If, however, the "direct" genetic effect (containing all effects on obesity not through gene expression) is separated from the indirect effects through gene expression, it can be revealed using the approach described in the following, and might be able to provide hypotheses about the regulatory mechanisms. In such situations, C-JAMP as well as regression models are often not able to separate direct and indirect effects, and can result in biased effects of the marginal parameters.

For a valid estimation and testing of the direct genetic - and generally molecular - effects, it is important to consider all potential effects between the genetic marker, phenotypes, and confounding factors. To this aim, under a directed acyclic graph setting, a new method is proposed to estimate and test the direct genetic effect on the primary phenotype. The standard error of the direct genetic effect estimate is estimated by using the so-called robust Huber-White sandwich variance estimator. Using unbiased estimation equations allows drawing on the known asymptotic properties of estimators and test statistics. The proposed approach called CIEE (causal inference method based on estimating equations) can be applied to different models with different error distributions, and it is presented here for the analysis of primary phenotypes that are normally-distributed (quantitative) or log-normally distributed time-to-event traits.

3.2.1 Statistical details

For disentangling direct and indirect genetic effects through intermediate phenotypes, causal diagrams (Pearl, 1995) are helpful for visualizing the research setting. Here, the DAG in Figure 3.4 is considered, which includes the direct effect of a genetic marker X on the primary phenotype Y and an indirect genetic effect through a secondary phenotype K. The model further includes measured and unmeasured factors L and U, respectively, which potentially confound the effect of K on Y. The goal of this study is to estimate and test the direct genetic effect α_{XY}, while removing the indirect effect of X on Y through K, and with robustness against effects of L and U. Without restriction of generality, it is assumed that there are no factors affecting X and that any factors are included as covariates in the analysis or have been dealt with using other approaches. For genetic effects, such factors are limited to family structure or population stratification, which can be included by methods described in Price et al. (2006) or Eu-ahsunthornwattana et al. (2014). For the investigation of other predictors X than genetic markers, previous knowledge has to be available to justify the assumed model. In the assumed DAG in Figure 3.4, it is assumed that $\alpha_{LY} = 0$ so that L is a factor influencing K. However, CIEE provides also valid inference if L is a measured confounder of $K \to Y$ (i.e., $\alpha_{LY} \neq 0$ and $\alpha_{XL} = 0$), as will be shown in section 4.3. If both $\alpha_{XL} \neq 0$ and $\alpha_{LY} \neq 0$, then the effect of L as intermediate phenotype can be removed from Y in the analysis analogously to K.

CIEE follows the general idea of the two-stage sequential G-estimation method (Vansteelandt et al., 2009), which first removes the effect of intermediate phenotypes from the primary phenotype, and then tests the genetic association with the adjusted primary phenotype. As a major difference, CIEE is one-stage and coefficient estimates of all parameters are obtained simultaneously by solving estimating equations.

For some background, consider observations of the primary phenotype $y_i, i = 1, \ldots, n$, and of covariates $z_{1i}, \ldots, z_{di}, i = 1, \ldots, n$.

Definition 3.2.1 (Estimating function, estimating equation)**.** An *estimating function* $\boldsymbol{U}(\boldsymbol{\omega})$ for a $(k \times 1)$ parameter vector $\boldsymbol{\omega}$ is a $(k \times 1)$ vector $\boldsymbol{U}(\boldsymbol{\omega}) = \sum U_i(\boldsymbol{\omega})$ of real-valued functions of $y_i, z_{1i}, \ldots, z_{di}, \boldsymbol{\omega}$. An *estimating equation* is an equation $\boldsymbol{U}(\boldsymbol{\omega}) = \mathbf{0}$ based on an estimating function. It is called *unbiased* if $E[\boldsymbol{U}_i(\boldsymbol{\omega})] = \mathbf{0}$.

Estimating equations can be used to obtain an estimate $\widehat{\boldsymbol{\omega}}$ of $\boldsymbol{\omega}$. For unbiased (and some asymptotically unbiased) estimating equations, asymptotic properties of $\widehat{\boldsymbol{\omega}}$ have been established that will be used for the estimation and inference in CIEE.

This also allows obtaining robust sandwich standard error estimates considering the additional variability of the estimates from the phenotype adjustment.

3.2.1.1 Analysis of a quantitative primary trait with CIEE

At first, I focus on the analysis of a quantitative, normally-distributed primary phenotype Y with n independent observations. For this situation, the sequential G-estimation method (Vansteelandt et al., 2009) is as follows. In the first stage, the effect of K on Y, α_1, is estimated and $\widehat{\alpha}_1$ is obtained using the LS estimation method under the model

$$Y_i = \alpha_0 + \alpha_1 k_i + \alpha_2 x_i + \alpha_3 l_i + \varepsilon_i, \;\; \varepsilon_i \sim N\left(0, \sigma_1^2\right). \tag{3.2.1}$$

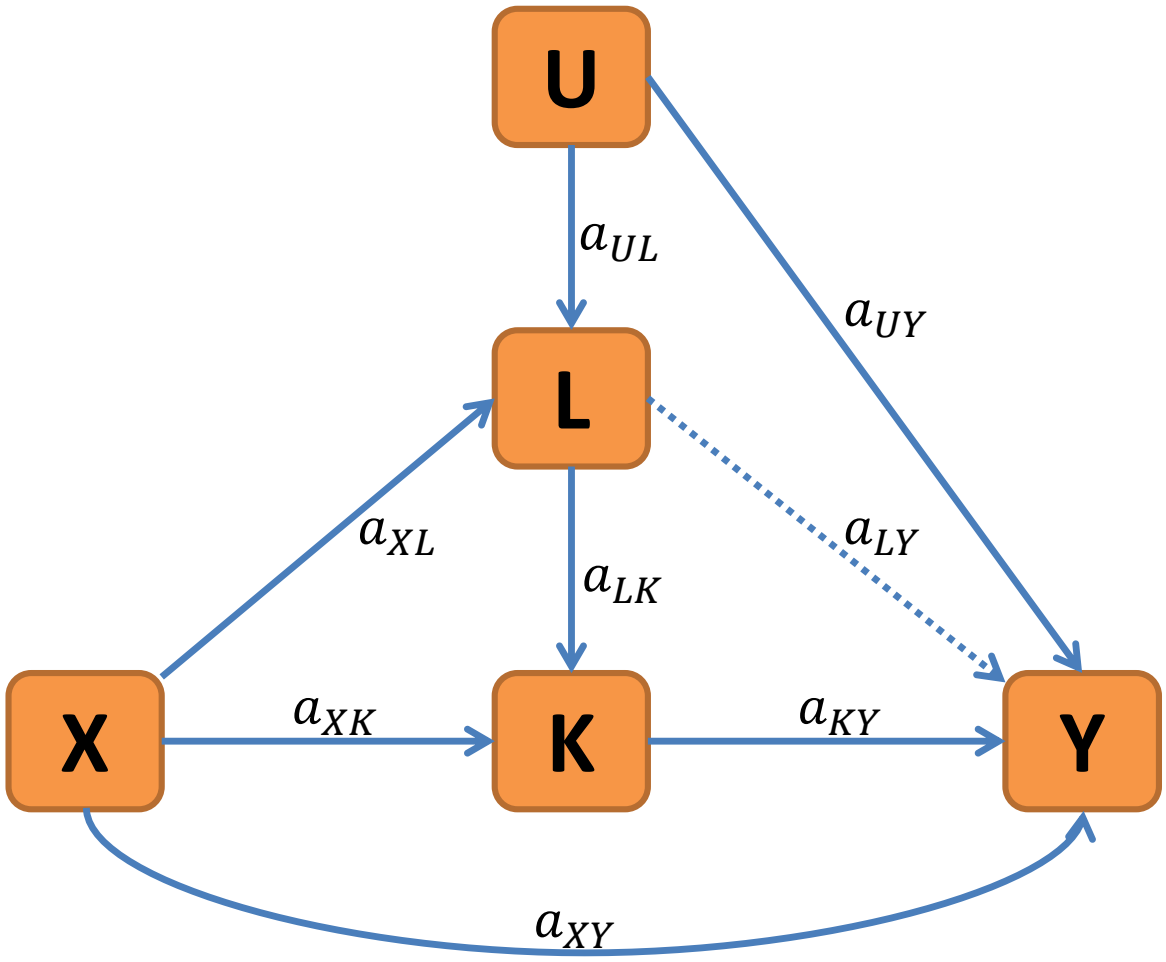


Figure 3.4: Assumed DAG in the methods description of CIEE. Y is the primary outcome measure of interest; K is a secondary phenotype; X is the genetic marker of interest and α_{XY} is the direct effect of interest. It is assumed that $\alpha_{LY} = 0$ so that L is a measured predictive factor of K, however CIEE is also valid if L is a measured confounder of $K \to Y$ (i.e., $\alpha_{XL} = 0$). U represents unmeasured factors and confounders potentially influencing L and Y.

Then, to block all indirect paths of X on the primary phenotype Y, the adjusted phenotype $\tilde{Y}$ is obtained by removing the effect of K on Y with

$$\widetilde{y_i} = y_i - \bar{y} - \widehat{\alpha}_1(k_i - \bar{k}), \tag{3.2.2}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{k} = \frac{1}{n}\sum_{i=1}^{n} k_i$. In the second stage, the significance of the direct effect of X on Y, α_{XY}, is tested under the model

$$\widetilde{Y_i} = \alpha_4 + \alpha_{XY} x_i + \varepsilon_i', \quad \varepsilon_i' \sim N\left(0, \sigma_2^2\right) \tag{3.2.3}$$

using the proposed test statistic in Vansteelandt and colleagues (2009). Building on this idea, I propose to formulate unbiased estimating equations $\boldsymbol{U}(\boldsymbol{\omega}) = \mathbf{0}$ are formulated for a consistent estimation of the unknown parameter vector $\boldsymbol{\omega} = \left(\alpha_0,\ \alpha_1,\ \alpha_2,\ \alpha_3,\ \sigma_1^2,\ \alpha_4,\ \alpha_{XY},\ \sigma_2^2\right)^T$, where

$$\boldsymbol{U}(\boldsymbol{\omega}) = \left(\frac{\partial l_1(\boldsymbol{\omega})}{\partial \alpha_0}, \frac{\partial l_1(\boldsymbol{\omega})}{\partial \alpha_1}, \frac{\partial l_1(\boldsymbol{\omega})}{\partial \alpha_2}, \frac{\partial l_1(\boldsymbol{\omega})}{\partial \alpha_3}, \frac{\partial l_1(\boldsymbol{\omega})}{\partial \sigma_1^2}, \frac{\partial l_2(\boldsymbol{\omega})}{\partial \alpha_4}, \frac{\partial l_2(\boldsymbol{\omega})}{\partial \alpha_{XY}}, \frac{\partial l_2(\boldsymbol{\omega})}{\partial \sigma_2^2} \right)^T \tag{3.2.4}$$

$$l_1(\boldsymbol{\omega}) = \sum_{i=1}^{n} \left[-\log(\sigma_1) + \log\left(\phi\left(\frac{y_i - \alpha_0 - \alpha_1 k_i - \alpha_2 x_i - \alpha_3 l_i}{\sigma_1} \right) \right) \right] \tag{3.2.5}$$

$$l_2(\boldsymbol{\omega}) = \sum_{i=1}^{n} \left[-\log(\sigma_2) + \log\left(\phi\left(\frac{y_i - \bar{y} - \alpha_1 (k_i - \bar{k}) - \alpha_4 - \alpha_{XY} x_i}{\sigma_2} \right) \right) \right], \tag{3.2.6}$$

and $\phi(.)$ is the PDF of the standard normal distribution. To give an intuition on how these estimating equations are obtained, $l_1(\boldsymbol{\omega})$ is the log-likelihood function under the model in (3.2.1) and $l_2(\boldsymbol{\omega})$ is the log-likelihood function under the model in (3.2.3) given that α_1 is known. Therefore, by solving the first five estimating equations based on $l_1(\boldsymbol{\omega})$ in equation (3.2.5), the model in (3.2.1) is fitted to obtain estimates of $\alpha_0, \alpha_1, \alpha_2, \alpha_3, \sigma_1^2$, that is obtaining the MLEs under the model in (3.2.1). Analogously, solving the last three estimating equations based on $l_2(\boldsymbol{\omega})$ in equation (3.2.6) yields estimates of $\alpha_4, \alpha_{XY}, \sigma_2^2$. Hence, the estimate of $\boldsymbol{\omega}$ is obtained, denoted by $\widehat{\boldsymbol{\omega}}$, by solving $\boldsymbol{U}(\boldsymbol{\omega}) = \boldsymbol{0}$. In more detail, the estimating equations can be derived as:

$$\begin{aligned}
U_1(\boldsymbol{\omega}) &= \frac{\partial l_1(\boldsymbol{\omega})}{\partial \alpha_0} \\
&= \frac{\partial \sum_{i=1}^{n} \left[-\log(\sigma_1) + \log\left(\phi\left(\frac{y_i - \alpha_0 - \alpha_1 k_i - \alpha_2 x_i - \alpha_3 l_i}{\sigma_1} \right) \right) \right]}{\partial \alpha_0} \\
&= \frac{\partial \left[-\frac{n}{2}\log(2\pi) - n\log(\sigma_1) - \frac{1}{2\sigma_1^2} \sum_{i=1}^{n} (y_i - \alpha_0 - \alpha_1 k_i - \alpha_2 x_i - \alpha_3 l_i)^2 \right]}{\partial \alpha_0} \\
&= \frac{n}{\sigma_1^2} \left(\bar{y} - \alpha_0 - \alpha_1 \bar{k} - \alpha_2 \bar{x} - \alpha_3 \bar{l} \right) = 0,
\end{aligned}$$

and similarly

$$\begin{aligned}
U_2(\boldsymbol{\omega}) &= \frac{\partial l_1(\boldsymbol{\omega})}{\partial \alpha_1} = \frac{n}{\sigma_1^2} \left(\overline{ky} - \alpha_0 \bar{k} - \alpha_1 \overline{k^2} - \alpha_2 \overline{kx} - \alpha_3 \overline{kl} \right) = 0, \\
U_3(\boldsymbol{\omega}) &= \frac{\partial l_1(\boldsymbol{\omega})}{\partial \alpha_2} = \frac{n}{\sigma_1^2} \left(\overline{xy} - \alpha_0 \bar{x} - \alpha_1 \overline{xk} - \alpha_2 \overline{x^2} - \alpha_3 \overline{xl} \right) = 0, \\
U_4(\boldsymbol{\omega}) &= \frac{\partial l_1(\boldsymbol{\omega})}{\partial \alpha_3} = \frac{n}{\sigma_1^2} \left(\overline{ly} - \alpha_0 \bar{l} - \alpha_1 \overline{lk} - \alpha_2 \overline{lx} - \alpha_3 \overline{l^2} \right) \quad = 0,
\end{aligned}$$

where $\bar{v}$ denotes the arithmetic mean of a variable v. This can be simplified to

$$\begin{pmatrix} U_1(\boldsymbol{\omega}) \\ U_2(\boldsymbol{\omega}) \\ U_3(\boldsymbol{\omega}) \\ U_4(\boldsymbol{\omega}) \end{pmatrix} = \frac{1}{\sigma_1^2} \boldsymbol{Z} \left(y - \boldsymbol{Z}^T \boldsymbol{\alpha} \right)^T = 0 \quad \text{with } \boldsymbol{Z} = (1, K, X, L)^T, \ \boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2, \alpha_3)^T,$$

and

$$\begin{aligned} U_5(\boldsymbol{\omega}) = & -\frac{n}{2\sigma_1^2} - \frac{1}{2\sigma_1^4}\left(y - \boldsymbol{Z}^T\boldsymbol{\alpha}\right)^T\left(y - \boldsymbol{Z}^T\boldsymbol{\alpha}\right) = 0, \\ \begin{pmatrix} U_6(\boldsymbol{\omega}) \\ U_7(\boldsymbol{\omega}) \end{pmatrix} = & \frac{1}{\sigma_1^2}\tilde{\boldsymbol{Z}}\left(\tilde{y} - \tilde{\boldsymbol{Z}}^T\tilde{\boldsymbol{\alpha}}\right)^T \quad \text{with } \tilde{\boldsymbol{Z}} = (1, X)^T,\ \tilde{\boldsymbol{\alpha}} = (\alpha_4, \alpha_{XY})^T, \\ U_8(\boldsymbol{\omega}) = & -\frac{n}{2\sigma_2^2} - \frac{1}{2\sigma_2^4}\left(\tilde{y} - \tilde{\boldsymbol{Z}}^T\tilde{\boldsymbol{\alpha}}\right)^T\left(\tilde{y} - \tilde{\boldsymbol{Z}}^T\tilde{\boldsymbol{\alpha}}\right) = 0. \end{aligned}$$

As a difference to the two-stage sequential G-estimation method, all parameters in $\boldsymbol{\omega}$ are estimated simultaneously and the additional variability obtained in the phenotype adjustment in (3.2.2) is considered by using the robust Huber-White sandwich estimator of the standard error of $\widehat{\boldsymbol{\omega}}$. The point estimate of the genetic effect $\widehat{\alpha}_{XY}$ from CIEE coincides in this model with the estimate from the two-stage sequential G-estimation method. Based on the Theorem 3.2.2 shown below, the robust Huber-White sandwich estimate of the standard error of $\widehat{\alpha}_{XY}$ can be obtained as $\widehat{SE}(\widehat{\alpha}_{XY}) = \sqrt{\frac{1}{n}\boldsymbol{C_n}(\widehat{\boldsymbol{\omega}})_{7,7}}$. Using the estimate of α_{XY} and its standard error estimate, a Wald-type test statistic can be computed for testing $H_0 : \alpha_{XY} = 0$ versus $H_A : \alpha_{XY} \neq 0$. Here, the large-sample Wald-type test statistic $W = \frac{\widehat{\alpha}_{XY}}{\widehat{SE}(\widehat{\alpha}_{XY})}$ is used which has asymptotically standard normal $N(0,1)$ distribution.

Theorem 3.2.2 (White, 1982)**.** *Let U_j be the j-th element of the vector in equation (3.2.4) and $q = 8$. Under mild regularity conditions, $\sqrt{n}(\widehat{\boldsymbol{\omega}} - \boldsymbol{\omega})$ is asymptotically normally distributed with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{C}(\boldsymbol{\omega})$, which can be consistently estimated with $\boldsymbol{C_n}(\widehat{\boldsymbol{\omega}})$, where*

$$\boldsymbol{C_n}(\boldsymbol{\omega}) = \boldsymbol{A_n}(\boldsymbol{\omega})^{-1}\,\boldsymbol{B_n}(\boldsymbol{\omega})\,\left[\boldsymbol{A_n}(\boldsymbol{\omega})^{-1}\right]^T \tag{3.2.7}$$

$$\boldsymbol{A_n}(\boldsymbol{\omega}) = -\frac{1}{n}\left(\frac{\partial U(\boldsymbol{\omega})}{\partial \boldsymbol{\omega}}\right) \tag{3.2.8}$$

$$\boldsymbol{B_n}(\boldsymbol{\omega}) = \frac{1}{n}\sum_{i=1}^{n}\left[U_j(y_i, k_i, x_i, l_i; \boldsymbol{\omega}) \cdot U_k(y_i, k_i, x_i, l_i; \boldsymbol{\omega})^T\right]_{j,k=1\ldots q} \tag{3.2.9}$$

3.2.1.2 Analysis of a time-to-event primary trait with CIEE

For the analysis of survival times T, the right-censoring scheme with observed survival times $t_i = \min(T_i, C_i)$ and censoring indicator $\delta_i = I[T_i \leq C_i]$ is considered for a random sample of individuals $i = 1, \ldots, n$, where T_i is the survival time, C_i is the censoring time and $I[.]$ is the indicator function. It is assumed that censoring is noninformative. I consider the AFT, or log-linear, model of the form

$$Y_i = \log(T_i) = \alpha_0 + \alpha_1 k_i + \alpha_2 x_i + \alpha_3 l_i + \sigma_1 \varepsilon_i, \quad \sigma_1 > 0 \tag{3.2.10}$$

for the phenotype adjustment. The error term in equation (3.2.10) can come from any distribution, and here the focus is on the log-linear model with $\varepsilon_i \sim N(0,1)$ for illustration. The estimating equations can be constructed similarly as described in the previous section, but in order to remove the effect of K from Y, the true underlying log-survival times Y_{est} need to be estimated for censored survival times. Y_{est} equals the observed log-survival time Y for uncensored survival times. To estimate Y_{est} for a censored survival time, the conditional

expectation of Y given that it is greater than the observed log-transformed right-censoring time and given the covariates is obtained (Konigorski et al., 2014), see Lemma 3.2.3. Then, the adjusted phenotype is computed using

$$\widetilde{y_i} = y_{est,i} - \overline{y_{est}} - \alpha_1(k_i - \bar{k}) \tag{3.2.11}$$

with $y_{est,i}$ obtained from equation (3.2.13) and $\overline{y_{est}} = \frac{1}{n}\sum_{i=1}^{n} y_{est,i}$. Finally, the direct genetic effect on the adjusted phenotype is modeled using

$$\widetilde{Y_i} = \alpha_4 + \alpha_{XY}x_i + \varepsilon_i', \;\; \varepsilon_i' \sim N\left(0, \sigma_2^2\right). \tag{3.2.12}$$

Lemma 3.2.3. *Consider the log-linear model in (3.2.10), let $f(.)$ denote the probability density function of Y given K, X, L and $F(.)$ the cumulative distribution function of Y given K, X, L under this model. Let further $\boldsymbol{\omega} = \left(\alpha_0, \alpha_1, \alpha_2, \alpha_3, \sigma_1, \alpha_4, \alpha_{XY}, \sigma_2^2\right)^T$, $\phi(.)$ denote the standard normal PDF, and $\Phi(.)$ the standard normal CDF. Then, the true underlying log-survival times Y_{est} are*

$$y_{est,i} = \delta_i \cdot y_i + (1-\delta_i) \cdot \left(\alpha_0 + \alpha_1 k_i + \alpha_2 x_i + \alpha_3 l_i + \frac{\sigma_1 \cdot \phi\left(\frac{y_i - \alpha_0 - \alpha_1 k_i - \alpha_2 x_i - \alpha_3 l_i}{\sigma_1}\right)}{1 - \Phi\left(\frac{y_i - \alpha_0 - \alpha_1 k_i - \alpha_2 x_i - \alpha_3 l_i}{\sigma_1}\right)} \right)$$

Proof Lemma 3.2.3:

$$\begin{aligned} y_{est,i} &= \delta_i \cdot y_i + (1-\delta_i) \cdot E\left[Y_i | Y_i > y_i, k_i, x_i, l_i\right] \\ &= \delta_i \cdot y_i + (1-\delta_i) \cdot \left(E\left[Y_i | k_i, x_i, l_i\right] + \frac{{\sigma_1}^2 \cdot f\left(y_i | k_i, x_i, l_i; \boldsymbol{\omega}\right)}{1 - F\left(y_i | k_i, x_i, l_i; \boldsymbol{\omega}\right)} \right) \\ &= \delta_i \cdot y_i + (1-\delta_i) \cdot \left(\alpha_0 + \alpha_1 k_i + \alpha_2 x_i + \alpha_3 l_i + \frac{\sigma_1 \cdot \phi\left(\frac{y_i - \alpha_0 - \alpha_1 k_i - \alpha_2 x_i - \alpha_3 l_i}{\sigma_1}\right)}{1 - \Phi\left(\frac{y_i - \alpha_0 - \alpha_1 k_i - \alpha_2 x_i - \alpha_3 l_i}{\sigma_1}\right)} \right). \end{aligned} \tag{3.2.13}$$

Here, with $\boldsymbol{z} = (k, x, l)^T$, $\nu = \frac{1}{2}\left(\frac{y-E[Y|\boldsymbol{z}]}{\sigma_1}\right)^2$, $d\nu = \frac{y-E[Y|\boldsymbol{z}]}{\sigma_1}dy$,

$$
\begin{aligned}
E\left[Y|Y>y_0,\boldsymbol{z}\right] &= \int_{y_0}^{\infty} y f\left(y|\boldsymbol{z};\boldsymbol{\omega}\right) dy \Big/ 1-F\left(y_0|\boldsymbol{z};\boldsymbol{\omega}\right) \\
&= \sigma_1 \int_{y_0}^{\infty} \frac{y}{\sigma_1} \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{1}{2}\left(\frac{y-E\left[Y|\boldsymbol{z}\right]}{\sigma_1}\right)^2\right) dy \Big/ 1-F\left(y_0|\boldsymbol{z};\boldsymbol{\omega}\right) \\
&= \frac{\left[\sigma_1 \int_{y_0}^{\infty} \frac{y-E[Y|\boldsymbol{z}]}{\sigma_1} \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{1}{2}\left(\frac{y-E[Y|\boldsymbol{z}]}{\sigma_1}\right)^2\right) dy + E\left[Y|\boldsymbol{z}\right] \int_{y_0}^{\infty} f\left(y|\boldsymbol{z};\boldsymbol{\omega}\right) dy\right]}{1-F\left(y_0|\boldsymbol{z};\boldsymbol{\omega}\right)} \\
&= \left[\sigma_1 \int_{\frac{1}{2}\left(\frac{y_0-E[Y|\boldsymbol{z}]}{\sigma_1}\right)^2}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\nu\right) d\nu + E\left[Y|\boldsymbol{z}\right]\left(1-F\left(y_0|\boldsymbol{z};\boldsymbol{\omega}\right)\right)\right] \Big/ 1-F\left(y_0|\boldsymbol{z};\boldsymbol{\omega}\right) \\
&= \left[\frac{\sigma_1}{\sqrt{2\pi}}\left(-\exp\left(-\nu\right)\right)\Big|_{\frac{1}{2}\left(\frac{y_0-E[Y|\boldsymbol{z}]}{\sigma_1}\right)^2}^{\infty} + E\left[Y|\boldsymbol{z}\right]\left(1-F\left(y_0|\boldsymbol{z};\boldsymbol{\omega}\right)\right)\right] \Big/ 1-F\left(y_0|\boldsymbol{z};\boldsymbol{\omega}\right) \\
&= \left[\frac{\sigma_1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{y_0-E\left[Y|\boldsymbol{z}\right]}{\sigma_1}\right)^2\right) + E\left[Y|\boldsymbol{z}\right]\left(1-F\left(y_0|\boldsymbol{z};\boldsymbol{\omega}\right)\right)\right] \Big/ 1-F\left(y_0|\boldsymbol{z};\boldsymbol{\omega}\right) \\
&= E\left[Y|\boldsymbol{z}\right] + \frac{\sigma_1^2 \cdot f\left(y|\boldsymbol{z};\boldsymbol{\omega}\right)}{1-F\left(y|\boldsymbol{z};\boldsymbol{\omega}\right)}.
\end{aligned}
$$

□

The estimating equations for estimating $\boldsymbol{\omega} = \left(\alpha_0, \alpha_1, \alpha_2, \alpha_3, \sigma_1, \alpha_4, \alpha_{XY}, \sigma_2^2\right)^T$ are

$$
\boldsymbol{U}\left(\boldsymbol{\omega}\right) = \left(\frac{\partial l_1\left(\boldsymbol{\omega}\right)}{\partial \alpha_0}, \frac{\partial l_1\left(\boldsymbol{\omega}\right)}{\partial \alpha_1}, \frac{\partial l_1\left(\boldsymbol{\omega}\right)}{\partial \alpha_2}, \frac{\partial l_1\left(\boldsymbol{\omega}\right)}{\partial \alpha_3}, \frac{\partial l_1\left(\boldsymbol{\omega}\right)}{\partial \sigma_1}, \frac{\partial l_2\left(\boldsymbol{\omega}\right)}{\partial \alpha_4}, \frac{\partial l_2\left(\boldsymbol{\omega}\right)}{\partial \alpha_{XY}}, \frac{\partial l_2\left(\boldsymbol{\omega}\right)}{\partial \sigma_2^2}\right)^T = \boldsymbol{0} \tag{3.2.14}
$$

with

$$
\begin{aligned}
l_1\left(\boldsymbol{\omega}\right) = \sum_{i=1}^{n} \Bigg[&-\delta_i \log\left(\sigma_1\right) + \delta_i \log\left(\phi\left(\frac{y_i-\alpha_0-\alpha_1 k_i-\alpha_2 x_i-\alpha_3 l_i}{\sigma_1}\right)\right) \\
&+ \left(1-\delta_i\right) \log\left(1-\Phi\left(\frac{y_i-\alpha_0-\alpha_1 k_i-\alpha_2 x_i-\alpha_3 l_i}{\sigma_1}\right)\right)\Bigg]
\end{aligned} \tag{3.2.15}
$$

and

$$
l_2\left(\boldsymbol{\omega}\right) = \sum_{i=1}^{n} \left[-\log\left(\sigma_2\right) + \log\left(\phi\left(\frac{y_{est,i}-\overline{y_{est}}-\alpha_1\left(k_i-\bar{k}\right)-\alpha_4-\alpha_{XY}x_i}{\sigma_2}\right)\right)\right], \tag{3.2.16}
$$

where $\phi(.)$ and $\Phi(.)$ are the standard normal PDF and CDF. $\boldsymbol{U}(\boldsymbol{\omega}) = \boldsymbol{0}$ are unbiased estimating equations with $\sqrt{n}\left(\widehat{\boldsymbol{\omega}}-\boldsymbol{\omega}\right) \xrightarrow{D} N\left(0, \boldsymbol{C}\left(\boldsymbol{\omega}\right)\right)$ where $\boldsymbol{C}(\boldsymbol{\omega})$ is estimated as described in the previous section. Here, $l_1(\boldsymbol{\omega})$ is the log-likelihood function under the model in equation (3.2.10) and $l_2(\boldsymbol{\omega})$ is the log-likelihood function under the model in equation (3.2.12) given that $\alpha_0, \alpha_1, \alpha_2, \alpha_3, \sigma_1$ are known. By solving the first five estimating equations based on $l_1(\boldsymbol{\omega})$ in equation (3.2.15), estimates of $\alpha_0, \alpha_1, \alpha_2, \alpha_3, \sigma_1$ are obtained and solving the last three estimating equations based on $l_2(\boldsymbol{\omega})$ in (3.2.16) yields estimates of $\alpha_4, \alpha_{XY}, \sigma_2^2$. In more detail,

the functions $l_1(\boldsymbol{\omega})$ and $l_2(\boldsymbol{\omega})$ are derived as follows:

$$\begin{aligned}
l_1(\boldsymbol{\omega}) &= \sum_{i=1}^{n} \delta_i \cdot \log\left(f\left(y_i|k_i, x_i, l_i; \boldsymbol{\omega}\right)\right) + (1-\delta_i) \cdot \log\left(1 - F\left(y_i|k_i, x_i, l_i; \boldsymbol{\omega}\right)\right) \\
&= \sum_{i=1}^{n} \left[\delta_i \log\left(\frac{1}{\sigma_1}\phi\left(\frac{y_i - \alpha_0 - \alpha_1 k_i - \alpha_2 x_i - \alpha_3 l_i}{\sigma_1}\right)\right) \right. \\
&\qquad \left. + (1-\delta_i)\log\left(1 - \Phi\left(\frac{y_i - \alpha_0 - \alpha_1 k_i - \alpha_2 x_i - \alpha_3 l_i}{\sigma_1}\right)\right)\right] \\
&= \sum_{i=1}^{n} \left[-\delta_i \log\left(\sigma_1\right) + \delta_i \log\left(\phi\left(\frac{y_i - \alpha_0 - \alpha_1 k_i - \alpha_2 x_i - \alpha_3 l_i}{\sigma_1}\right)\right) \right. \\
&\qquad \left. + (1-\delta_i)\log\left(1 - \Phi\left(\frac{y_i - \alpha_0 - \alpha_1 k_i - \alpha_2 x_i - \alpha_3 l_i}{\sigma_1}\right)\right)\right],
\end{aligned}$$

$$\begin{aligned}
l_2(\boldsymbol{\omega}) &= \sum_{i=1}^{n} \log\left(\frac{1}{\sigma_2}\phi\left(\frac{y_{est,i} - \overline{y_{est}} - \alpha_1\left(k_i - \bar{k}\right) - \alpha_4 - \alpha_{XY}x_i}{\sigma_2}\right)\right) \\
&= \sum_{i=1}^{n} \left[-\log\left(\sigma_2\right) + \log\left(\phi\left(\frac{y_{est,i} - \overline{y_{est}} - \alpha_1\left(k_i - \bar{k}\right) - \alpha_4 - \alpha_{XY}x_i}{\sigma_2}\right)\right)\right] \\
&= \sum_{i=1}^{n} \left[-\log\left(\sigma_2\right) + \log\left(\phi\left(\frac{\delta_i y_i + (1-\delta_i)\left(\alpha_0 + \alpha_1 k_i + \alpha_2 x_i + \alpha_3 l_i + \frac{\sigma_1 \phi\left(\frac{y_i - \alpha_0 - \alpha_1 k_i - \alpha_2 x_i - \alpha_3 l_i}{\sigma_1}\right)}{1 - \Phi\left(\frac{y_i - \alpha_0 - \alpha_1 k_i - \alpha_2 x_i - \alpha_3 l_i}{\sigma_1}\right)}\right)}{\sigma_2} \right.\right.\right. \\
&\qquad \left.\left.\left. - \frac{\overline{y_{est}} - \alpha_1\left(k_i - \bar{k}\right) - \alpha_4 - \alpha_{XY}x_i}{\sigma_2}\right)\right)\right].
\end{aligned}$$

3.2.1.3 Estimation of standard errors using nonparametric bootstrap

As an alternative to the sandwich variance estimator based on estimating equations, nonparametric bootstrap (Efron, 1981) can be used to obtain the standard error estimate of the estimated direct genetic effect $\widehat{\alpha}_{XY}$. In step 1, a sample of n individuals is randomly selected from the data with replacement. In step 2, the point estimate $\widehat{\alpha}_{XY,t}$ is obtained by solving the estimating equations in (3.2.4) or (3.2.14). These two steps are performed B times and the bootstrap standard error estimate of $\widehat{\alpha}_{XY}$ can be obtained as the standard deviation of the $\widehat{\alpha}_{XY,t}, t = 1, \ldots, B$.

3.2.2 Computational details & implementation

CIEE is implemented in the R package `CIEE` for both the analysis of continuous and time-to-event traits subject to censoring as primary outcomes for the model in Figure 3.4. In addition, functions are available to fit the same DAG with traditional regression approaches and the structural equations modeling method. For an extension to include more covariates or intermediate phenotypes, the estimating equations and R functions have to be adapted.

`CIEE` is currently available upon request as a bundled package. It can be installed through the zipped tar file with

```
install.packages("pathtofile/CIEE_0.1.0.tar.gz", repos = NULL)
```

and then loaded with

```
library(CIEE)
```

CIEE was generated using RStudio Version 1.0.143 based on R version 3.4.1 (R Core Team, 2017) using the `devtools` package. The documentation of the functions and generation of help pages was done using the `roxygen2` package. Finally, `knitr` and `rmarkdown` were used for the detailed long-form documentation (vignette), which is available through

```
browseVignettes("CIEE")
```

More details for each function can be obtained through the regular help files, for example,

```
?ciee
```

In the following, an outline of the implementation is given, as well as details regarding the application to genome-wide analyses. Examples for the use of the R functions are shown in the appendix in section A.7.2.

3.2.2.1 Data generation functions

Simulated data can be generated from the model in Figure 3.4 using the `generate_data()` function, which generates data for the quantitative outcome Y (or for time-to-event outcomes T, $Y = \log(T)$ and censoring indicator C), intermediate phenotype K, a genetic marker X (SNV coded 0, 1, 2), and observed as well as unobserved confounders L, U.

3.2.2.2 Traditional regression and SEM functions

Two traditional approaches for fitting the model in Figure 3.4 are the multiple regression (MR) and regression of residuals (RR) methods:

MR: Obtain the LS estimate of α_{XY} by fitting

$$Y_i = \alpha_0 + \alpha_{XY} x_i + \alpha_1 k_i + \alpha_2 l_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma_1^2). \tag{3.2.17}$$

RR: First, obtain residuals $\widehat{\varepsilon}_{1i} = y_i - (\widehat{\alpha}_0 + \widehat{\alpha}_1 k_i + \widehat{\alpha}_2 l_i)$ by fitting

$$Y_i = \alpha_0 + \alpha_1 k_i + \alpha_2 l_i + \varepsilon_{1i}, \quad \varepsilon_{1i} \sim N(0, \sigma_1^2) \tag{3.2.18}$$

using the LS estimation. Second, obtain the LS estimate of α_{XY} by fitting

$$\widehat{\varepsilon}_{1i} = \alpha_3 + \alpha_{XY} x_i + \varepsilon_{2i}, \quad \varepsilon_{2i} \sim N(0, \sigma_2^2). \tag{3.2.19}$$

The two approaches are implemented in the functions `mult_reg()` and `res_reg()`. For the analysis of a quantitative primary trait, point estimates and standard error estimates of

the parameters as well as p-values from the default t-test of $H_0 : \alpha_{XY} = 0$ versus $H_A : \alpha_{XY} \neq 0$ based on the `lm()` function are obtained. Under the AFT setting, only the MR approach is implemented to fit the censored log-linear regression model in equation (3.2.10) using the `survreg()` function in the `survival` R package in order to obtain parameter estimates and to perform a Wald test for testing the null hypothesis $H_0 : \alpha_{XY} = 0$.

As another approach, the SEM method (Bollen, 1989) can be used to obtain estimates of α_{XY} and its standard error. The function `sem_appl()` can be used to apply the SEM method to the DAG in Figure 3.4 based on the following model equations:

$$L_i = \alpha_0 + \alpha_1 x_i + \varepsilon_{1i}, \qquad \varepsilon_{1i} \sim N\left(0, \sigma_1^2\right), \tag{3.2.20}$$

$$K_i = \alpha_2 + \alpha_3 x_i + \alpha_4 l_i + \varepsilon_{2i}, \qquad \varepsilon_{2i} \sim N\left(0, \sigma_2^2\right), \tag{3.2.21}$$

$$Y_i = \alpha_5 + \alpha_6 k_i + \alpha_{XY} x_i + \varepsilon_{3i}, \quad \varepsilon_{3i} \sim N\left(0, \sigma_3^2\right). \tag{3.2.22}$$

`sem_appl()` uses the `sem()` function in the `lavaan` R package (Rosseel, 2012) with default settings, to obtain parameter and standard error estimates and to obtain p-values for testing $H_0 : \alpha_{XY} = 0$ versus $H_A : \alpha_{XY} \neq 0$ using the default Wald-type test.

3.2.2.3 CIEE functions

Regarding the implementation of CIEE in this package, the `est_funct_expr()` function contains the main part of the l_1 and l_2 functions in (3.2.5)-(3.2.6) and (3.2.15)-(3.2.16) as an expression. The `get_estimates()` function obtains estimates of $\boldsymbol{\omega}$ in (3.2.1)-(3.2.3) (for normally-distributed Y) or (3.2.10)-(3.2.12) (for time-to-event Y) by using the `lm()` and `survreg()` functions for computational purposes. These estimates are identical to estimates obtained by solving the estimating equations.

In order to compute the robust Huber-White sandwich estimator of the parameters, in a first step, the `deriv_obj()` function computes the expression of all first and second derivatives of l_1 and l_2 with respect to $\alpha_0, \alpha_1, \alpha_2, \alpha_3, \sigma_1^2, \alpha_4, \alpha_{XY}, \sigma_2^2$ by using the `deriv()` function and the expressions from the `est_funct_expr()` function as input. Then, the `scores()` and `hessian()` functions are written to obtain the numerical values of all first and second derivatives for the observed data and parameter estimates.

With these derivations, the robust Huber-White sandwich estimator of the standard error can be obtained using the `sandwich_se()` function. Alternatively, bootstrap standard error estimates can be obtained using the `bootstrap_se()` function. Also, for comparison, the function `naive_se()` computes naive standard error estimates of $\widehat{\alpha}_0, \widehat{\alpha}_1, \widehat{\alpha}_2, \widehat{\alpha}_3, \widehat{\alpha}_4, \widehat{\alpha}_{XY}$ without accounting for the additional variability due to the two stages in (3.2.1)-(3.2.3) or (3.2.10)-(3.2.12).

Finally, the `ciee()` function allows an easy integrated use of all above functions. `ciee()` fits the model in equations (3.2.1)-(3.2.3) or (3.2.10)-(3.2.12) (e.g., the model in Figure 3.4), yields parameter estimates and standard error estimates, and performs hypothesis tests of all parameters, for CIEE, the traditional regression approaches, and the SEM method. `ciee_loop()` provides an extension of `ciee()` and allows the input of multiple exposure variables (e.g., multiple SNPs) to be tested sequentially. In the output of `ciee_loop()`, only the coefficient estimates, standard error estimates, and p-values with respect to the direct effect α_{XY} are provided. Both `ciee()` and `ciee_loop()` return `ciee` objects as output, so that

the written `summary.ciee()` function can be used through the generic `summary()` function to provide a reader-friendly output of the results.

3.2.2.4 CIEE application details

For a genome-wide association analysis of direct genetic effects, CIEE can be used by either applying the `ciee()` sequentially on all SNVs of interest, or by supplying a matrix of multiple SNVs to the `ciee_loop()` function. The computation times of both functions show little differences in general. For a higher number of supplied SNVs, the `ciee_loop()` function is slightly more efficient, but at the same time, it requires more memory when the SNVs matrix becomes very large. An overview about the computation time for the estimation and testing of one SNV is shown in Figure 3.5 for the different approaches, for different sample sizes for the analysis of a quantitative and time-to-event primary trait. It can be seen that while CIEE is slower compared to all other approaches, is is still very fast and takes less than 1s for a sample size of $n = 100,000$. The running time for analyzing a time-to-event trait is not affected by the amount of censoring and takes slightly longer.

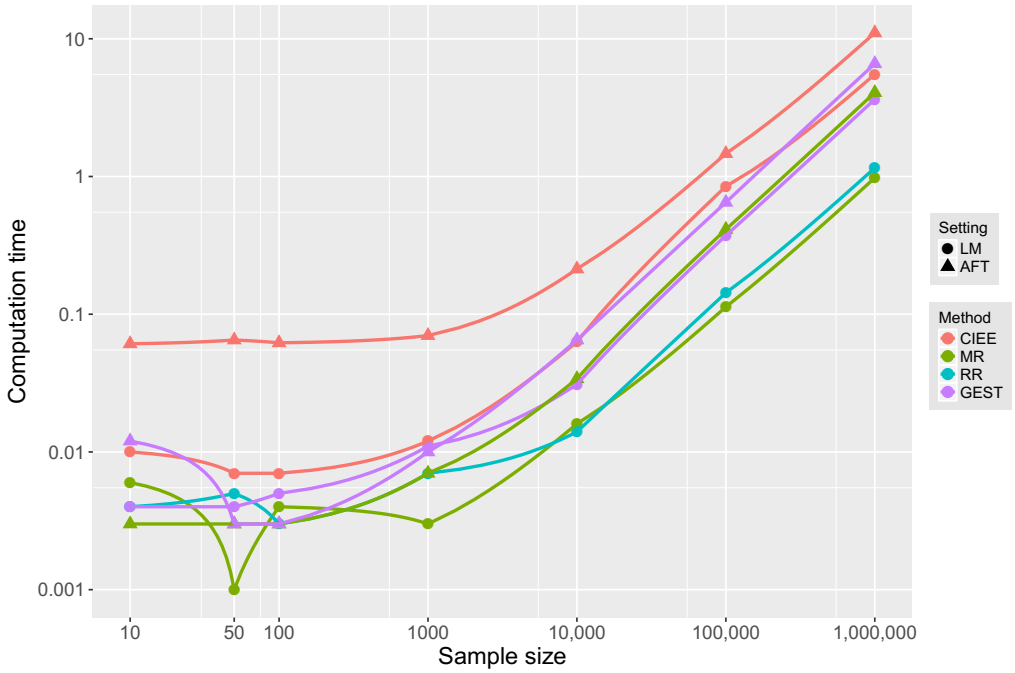


Figure 3.5: Computation time of CIEE. Shown are the mean computation times in seconds ('user times' obtained through the R `system.time()` function) based on 10 runs of the different approaches using the `ciee()` R function, for data generated in the power investigation in scenario 5 in Table 4.7, for the LM and AFT setting, for a SNV of MAF$= 0.2$, $a_{XY} = 0.1$ (and with 30% censoring).

A more detailed overview of the computation time and memory use of the different parts of the `ciee()` function are shown in Figures 3.6-3.7. They show that the computation time is quite evenly distributed over the multiple functions, and most of the running time is spent on deriving the first and second derivatives and obtaining the sandwich standard error estimate.

Flame Graph | Data | Options

Code	File	Memory (MB)		Time (ms)
▼ ciee	<expr>	-459.1	652.2	820
sandwich_se		-273.9	280.8	280
hessian		-34.3	67.9	110
scores		-22.5	0	10
▶ deriv_obj		-128.4	224.6	220
▶ get_estimates		0	70.2	190
▶ [		0	8.6	10

Sample Interval: 10ms 830ms

Figure 3.6: Detailed computation time and memory use of the `ciee()` R function under the LM setting. Shown are the memory use (in MB) and computation time (in ms) of each part of the `ciee()` R function, computed using the `profvis()` function of the `profvis` R package, for data generated in the power investigation in scenario 5 in Table 4.7 under the LM setting with $n = 100,000$, a SNV of MAF$= 0.2$ and $a_{XY} = 0.1$.

Flame Graph | Data | Options

Code	File	Memory (MB)		Time (ms)
▼ ciee	<expr>	-820.2	1113.1	1480
sandwich_se		-239.2	345.2	210
hessian		-120.2	82.4	120
▶ deriv_obj		-295.2	398.3	590
▶ get_estimates		-165.6	277.0	550
▶ [		0	10.2	10

Sample Interval: 10ms 1490ms

Figure 3.7: Detailed computation time and memory use of the `ciee()` R function under the AFT setting. Shown are the memory use (in MB) and computation time (in ms) of each part of the `ciee()` R function, computed using the `profvis()` function of the `profvis` R package, for data generated in the power investigation in scenario 5 in Table 4.7 under the AFT setting with $n = 100,000$, a SNV of MAF$= 0.2$, $a_{XY} = 0.1$ and with 30% censoring.

Chapter 4

Evaluation: Simulation studies

4.1 Existing single-marker tests versus multi-marker tests

In a first empirical investigation through simulation studies, an SMT (Wald-type t-test) and popular MMTs (a burden test, SKAT, and SKAT-O) were compared in genetic association studies of rare variants with a quantitative trait. A normally-distributed quantitative trait Y was generated given causal SNVs (under the alternative hypothesis scenarios) and two covariates, with different scenarios varying the percentage of causal variants, their effect sizes, and direction of effects. At first, the statistical properties of the SMT were assessed, and whether it provides valid estimation and inference. Then, all tests were compared regarding their empirical type I error and power to identify a causal gene.

4.1.1 Material and methods

4.1.1.1 Genetic data generation and simulation study set-up

For the main study, the genetic dataset provided in the `SKAT` package in R (Lee et al., 2016) was used, which contains 10,000 haplotypes over a 200kb region (including 3,845 SNVs) generated from a calibration coalescent model mimicking the linkage disequilibrium (LD) structure of European ancestry. This was chosen to make our study comparable to the evaluation of SKAT-O in (Lee et al., 2012). Accordingly, the kernels and other SKAT options were chosen to reach an optimal performance for the power of SKAT and SKAT-O.

Similar to (Lee et al., 2012), 10,000 3kb regions from the 200kb region were randomly sampled, to obtain genes with average length. Then, 2,000 haplotypes of these 10,000 genes were randomly paired to generate $m = 10,000$ replicates with genotypes of $n = 1,000$ individuals for the simulation study. With this, there were on average 33 non-monomorphic SNVs X_j in each replicate (min = 19; max = 52 SNVs), and in total 325,393 SNVs in all $m = 10,000$ replicates combined. Most of the SNVs were rare with MAF≤ 0.03. In detail, of the 325,393 SNVs, 132,797 SNVs had MAF = 0.0005 (minor allele count MAC = 1); 41,341 SNVs had MAF = 0.001 (MAC = 2); 50,836 SNVs had $0.001 <$ MAF ≤ 0.005 ($2 <$ MAC ≤ 10); 39,835 SNVs had $0.005 <$ MAF ≤ 0.03; and 60,584 SNVs had MAF > 0.03.

4.1.1.2 Phenotype generation

To evaluate the type I error and the power of the test statistics, Y was generated conditional on a binary covariate Z_1 and a normally distributed covariate Z_2 (and conditional on the causal SNVs X_j under the alternative hypotheses). The same model and parameter values were used as described in Lee et al. (2012) with additive genetic effects:

$$Y = 0.5z_1 + 0.5z_2 + \sum \beta_j x_j + \varepsilon \tag{4.1.1}$$

with $Z_1 \sim \text{Bin}\,(p = 0.5)\,,\ Z_2 \sim N\,(0,1)\,,\ \varepsilon \sim N\,(0,1)\,, \beta_j = c \cdot |\log_{10}\,(\text{MAF}_j)|$, where different values of c were considered in the simulation study. For the evaluation of type I error rates, $m = 10,000,000$ replicates were analyzed, and for the power comparisons, $m = 10,000$ replicates were used. The sample size of $n = 1,000$ individuals was chosen for all simulations. Table 4.1 gives an overview of the parameters in each scenario.

Investigation	Scenario	% of causal variants	Effect size weights	Direction of effect: Positive / Negative	Median (MAD) of explained variance in %
Type I error	0	0%	c=0	–	–
Power	1	5%	c=0.6	100% / 0%	0.9% (0.6)
	2	5%	c=0.3	100% / 0%	0.2% (0.2)
	3	5%	c=0.2	100% / 0%	0.1% (0.1)
	4	10%	c=0.6	100% / 0%	1.9% (1.4)
	5	10%	c=0.3	100% / 0%	0.5% (0.3)
	6	10%	c=0.2	100% / 0%	0.2% (0.2)
	7	20%	c=0.6	100% / 0%	3.8% (2.1)
	8	20%	c=0.3	100% / 0%	1.0% (0.6)
	9	20%	c=0.2	100% / 0%	0.4% (0.2)
	10	50%	c=0.6	100% / 0%	9.1% (3.0)
	11	50%	c=0.3	100% / 0%	2.4% (0.9)
	12	50%	c=0.2	100% / 0%	1.1% (0.4)
	13-24	as in scenarios 1-12		80% / 20%	as in scenarios 1-12
	25-36	as in scenarios 1-12		50% / 50%	as in scenarios 1-12

Table 4.1: Overview of the scenarios considered for the simulation study of the SMT and MMTs to evaluate their type I error and power. The scenarios vary the percentage of causal variants, their effect size, and the percentage of causal variants with effects in positive/ negative direction. Scenarios 13-24 and 25-36 have the same percentage of causal SNVs and the same effect sizes as scenarios 1-12, but 80% / 20% and 50% / 50% of effects in positive / negative direction. The percentage of causal rare variants is with respect to the total number of rare variants with $\text{MAF}_j \leq 0.03$ in the gene. The effect size of a variant with a given MAF on the trait Y is $\beta_j = c \cdot |\log_{10}(\text{MAF}_j)|$. The percentage of explained variance for a given gene is calculated as the sum of $2 \cdot \text{MAF}_j \cdot (1 - \text{MAF}_j) \cdot \beta_j^2 / Var\,(Y)$ over all variants X_j in the gene. Reported are the median and the median absolute deviation (MAD) of this heritability estimate over the 10,000 replicates.

For the type I error evaluation, empirical gene-level estimates were obtained for all approaches under the null hypothesis given in scenario 0 in Table 4.1. For the power evaluation, 36 different scenarios were considered, extending the investigation in Lee et al. (2012). They

differ in percentages of causal rare variants, effect sizes, and direction of the effects of causal rare variants. Causal variants were randomly chosen among all rare variants with MAF$\leq$ 0.03. Between 5% and 50% of the rare variants in a gene were set to be causal, and at least 1 rare variant per gene was chosen as causal under the alternative hypothesis. The median of the explained phenotypic variation by all causal SNVs in a gene was less than or equal to 1% in most scenarios, and ranged between 0.1% in scenario 3 (5% causal SNVs in a gene with small effect sizes) and 9.1% in scenario 10 (50% causal SNVs in a gene with larger effect sizes), calculated as the sum of $2 \cdot \text{MAF}_j \cdot (1 - \text{MAF}_j) \cdot \beta_j^2 / Var\,(Y)$ over all variants X_j in the gene (Laird & Lange 2011).

4.1.1.3 Investigated rare variant tests

In the comparison of the SMT and MMTs, all approaches were evaluated in their power to identify a causal gene. In a simulation study including m replicates of a gene with multiple SNVs, gene-level power estimates can be generally obtained with

$$\begin{aligned}\widehat{\text{Power}} &= \widehat{P}(\text{Reject } H_0 \mid H_A \text{ is true}) \\ &= \frac{\text{number of significant tests among the } m \text{ replicates generated from } H_A}{m}.\end{aligned}$$

This was done for MMTs, and for SMTs it becomes

$$\widehat{\text{Power}} = \frac{\text{number of replicates including at least one significant SNV}}{m}$$

and an appropriate multiple testing correction for the tests within the gene is used. This power calculation amounts to the *minP* approach (e.g., Madsen & Browning 2009; Basu & Pan 2011; Pan et al., 2014; Xing et al., 2012). To evaluate the performance of SMTs, a linear regression model of Y,

$$Y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_{XY} x_j + \varepsilon, \varepsilon \sim\ N(0, \sigma^2), \tag{4.1.2}$$

was separately fitted for each SNV X_j in a gene using the `lm()` function in R, which provides the MLE $\widehat{\beta}_{XY}$ of the coefficient β_{XY} and its standard error estimate $\widehat{SE}(\widehat{\beta}_{XY})$ using the unbiased estimate of σ^2. Then, the Wald-type t-test statistic $\frac{\widehat{\beta}_{XY}}{\widehat{SE}(\widehat{\beta}_{XY})} \sim t_{n-4}$ was obtained for testing the null hypothesis in (2.2.1). In order to account for testing multiple SNVs in a gene in the SMT to obtain gene-level tests, the Bonferroni and Benjamini-Hochberg (BH) correction (Benjamini & Hochberg, 1995) was applied in the type I error evaluation, and the BH-correction in the power estimation. With respect to multi-marker approaches, three different tests were conducted incorporating all SNVs in a replicate (i.e., gene): a burden test, SKAT, and SKAT-O. The burden test was conducted by obtaining a gene score as the summation of the minor alleles of all SNVs in the gene (i.e., $\sum x_j$) for each individual, and testing the null hypothesis given in (2.2.2) under the linear regression model

$$Y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_{XY} \sum x_j + \varepsilon \tag{4.1.3}$$

using the Wald-type t-test statistic obtained from the `lm()` function in R.

The sequence-kernel association test (SKAT; Wu et al., 2011) is a variance-component test of the null hypothesis in (2.2.2) based on the test statistic

$$Q_{SKAT} = (y - \widehat{y})^T \boldsymbol{K} (y - \widehat{y}), \tag{4.1.4}$$

where $\boldsymbol{K} = \boldsymbol{XWWX}^T$ is a weighted linear kernel matrix measuring the genetic similarity between individuals, $\boldsymbol{X} = (X_1, \ldots, X_k)$ is the $(n \times k)$ matrix of the k SNVs, $\boldsymbol{W} = \text{diag}(w_1, \ldots, w_k)$ is a $(k \times k)$ weight matrix with weights w_j based on the Beta distribution, $w_j \sim \text{Beta}\,(\text{MAF}(X_j), 1, 25)$ (see Supplementary Figure A.1 for an illustration), and $\widehat{y}$ are the predicted phenotypes based on the linear mixed model

$$Y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \boldsymbol{\beta_{XY}^T X} + \varepsilon \tag{4.1.5}$$

with $\boldsymbol{\beta_{XY}} = (\beta_{XY,1}, \ldots, \beta_{XY,k})^T$ following a multivariate distribution with exchangeable correlation structure. The optimal sequence-kernel association test (SKAT-O; Lee et al., 2012) is an extension of SKAT, deriving the variance-component test statistic

$$Q_\rho = (1 - \rho)\, Q_{SKAT} + \rho\, Q_{BURDEN}$$

as the optimal combination of burden and variance-component tests to maximize the power for testing the null hypothesis in (2.2.2). Here, ρ is determined through a data-adaptive grid search. Both SKAT and SKAT-O were computed using the default linear-weighted kernel (relating the genotypes to the phenotype) in the test statistic as it returned the highest power estimates among all possible options provided in the `SKAT()` function in the `SKAT` package in R. The weights in the weighted kernels of SKAT and SKAT-O are a function of the MAF of SNVs. For the p-value estimation, the default 'davies' setting was used for SKAT and 'optimal.adj' for SKAT-O. All common and rare SNVs in the gene were included in the analysis of the SMT and each MMT for a fair comparison. Additional results from analyzing rare variants only are provided in the appendix as a sensitivity analysis. All computations and visualizations were performed in R 2.15.0 and higher.

4.1.2 Results

4.1.2.1 Validity of estimates and test statistics of single-marker analysis

First, the coefficient estimates of SNVs and their standard error estimates in the linear regression model (4.1.2) were assessed under the null hypothesis (scenario 0 in Table 4.1) by focusing on singletons only, doubletons only, and all rare SNVs. The results showed that the effect estimates are unbiased and that the standard error estimates are equal to the standard deviations of the effect estimates, even when singletons and doubletons are under consideration (Table 4.2).

Variant type	$\text{mean}\left(\widehat{\beta}_{XY}\right)$	$\text{SD}\left(\widehat{\beta}_{XY}\right)$	$\text{mean}\left(\widehat{SE}_{\widehat{\beta}_{XY}}\right)$
Singleton	1.4×10^{-4}	1.00	1.00
Doubleton	1.0×10^{-4}	0.71	0.71
SNV with MAC=10	1.0×10^{-4}	0.32	0.32

Table 4.2: Bias and variance of the (restricted) MLEs from the SMT (linear regression) under the null hypothesis. Datasets were generated from the null model described in scenario 0 in Table 4.1 with size $n = 1,000$ for $m = 10,000,000$ replicates. The table shows the mean and standard deviation of the SNV effect estimates and the mean of the corresponding standard error estimates, for the 132,797,000 singletons in all replicates, the 41,341,000 doubletons in all replicates, and the 2,985,000 SNVs with 10 observed minor alleles (MAC = 10).

Investigation of the single-marker t-test statistic for testing the hypotheses in (2.2.1) showed that the distribution assumption holds for SNVs of all MAF (Figure 4.1, Supplementary Figure A.2).

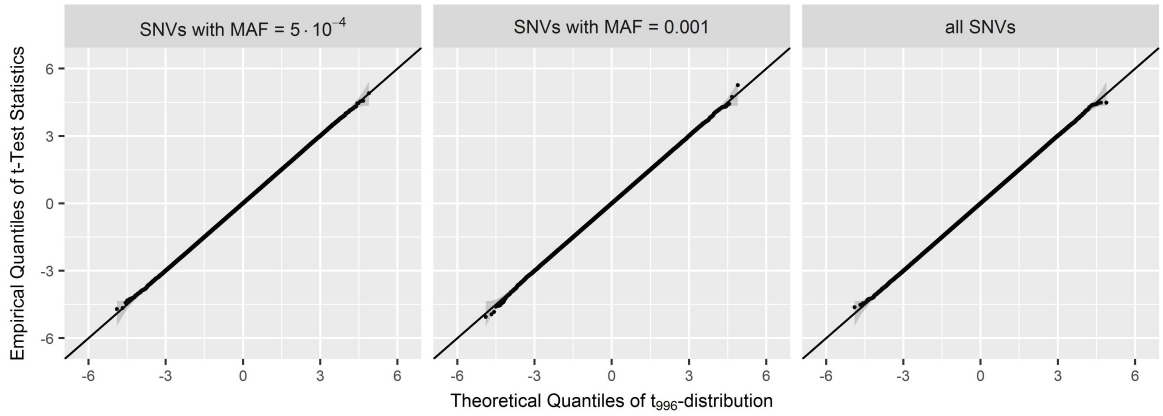


Figure 4.1: Distribution of SMT test statistics, for all SNVs, singletons only, and doubletons only. Datasets were generated from the null model described in scenario 0 in Table 4.1 of size $n = 1,000$ for $m = 10,000,000$ replicates. Q-Q plots are shown comparing the empirical quantiles of the t-test statistics of singletons (left panel), doubletons (middle panel), and all SNVs (right panel) to the theoretical quantiles of the $t_{df=1000-4}$ distribution. For computational purposes, each plot is based on a random sample of 1,000,000 t-test statistics, out of the 132,797,000 t-test statistics of all singletons in all replicates, out of the 41,341,000 t-test statistics of all doubletons in all replicates, and out of the 325,393,000 t-test statistics of all SNVs in all replicates. In grey ribbons, approximate 95% point-wise confidence intervals are shown.

4.1.2.2 Evaluation of type I error rates of the SMT and MMTs

Next, the empirical type I errors of the SMT and MMTs for testing the hypothesis in (2.2.2) were investigated for different nominal levels, to assess whether they are well-controlled (Table 4.3). The results indicate that empirical type I errors of all approaches are close to the nominal levels, allowing a valid comparison of the power estimates. For the power evaluations in the following, the BH correction was used for the SMT in order to account for the multiple testing of all SNVs within a gene and obtain gene-level power estimates, since some dependencies were observed between SNVs and since the empirical type I errors were closer to the nominal level compared to using the Bonferroni correction for most of the nominal levels.

Nominal α	MMTs			SMT	
	SKAT	SKAT-O	Burden	Bonferroni correction	BH correction
5×10^{-2}	4.94×10^{-2}	5.22×10^{-2}	5.01×10^{-2}	4.39×10^{-2}	4.93×10^{-2}
1×10^{-2}	0.97×10^{-2}	1.10×10^{-2}	1.00×10^{-2}	0.90×10^{-2}	0.99×10^{-2}
1×10^{-3}	0.94×10^{-3}	1.12×10^{-3}	1.00×10^{-3}	0.88×10^{-3}	0.97×10^{-3}
1×10^{-4}	0.92×10^{-4}	1.17×10^{-4}	1.03×10^{-4}	0.91×10^{-4}	0.98×10^{-4}
1×10^{-5}	0.92×10^{-5}	1.10×10^{-5}	1.12×10^{-5}	1.16×10^{-5}	1.25×10^{-5}
2.5×10^{-6}	2.40×10^{-6}	2.40×10^{-6}	3.50×10^{-6}	2.40×10^{-6}	2.90×10^{-6}

Table 4.3: Empirical type I error of the SMT and MMTs for different nominal α levels. Data was generated from the null model with size $n = 1,000$ for $m = 10,000,000$ replicates.

As a sensitivity check, misspecified distributions for the error term in model (4.1.2) were considered. Data was generated from the null model in (4.1.1) with error terms from the t-distribution with 4 and 8 degrees of freedom and from the standard log-normal distribution. Then, the SMT, SKAT, SKAT-O and burden tests were conducted to test the absence of genetic effects assuming normally-distributed errors as described in section 4.1.1. The results are shown in Supplementary Table A.1) and indicate that only the burden test has valid type I errors. The SKAT, SKAT-O and SMT are all sensitive to a misspecification of the error distribution, and the SMT shows a much higher inflation of empirical type I errors. The inflation can be slightly decreased by excluding SNVs with only 1 or 2 observed minor alleles but is still substantial and much higher than of SKAT and SKAT-O.

4.1.2.3 Power for identifying a causal gene with the SMT and MMTs

Figure 4.2 and Supplementary Tables A.2-A.3 show the power estimates for all test statistics under the scenarios 1-36 described in Table 4.1, when the type I error was 0.05 or 2.5×10^{-6}. The burden test had the lowest power among all tests in all scenarios when all rare and (non-causal) common variants were included in the analysis. Regarding a comparison among the other MMTs, SKAT-O had generally similar power with SKAT and the only noticeable differences were in scenarios 11, 12 (50% causal rare SNVs, all effects in the same direction) when the power of SKAT-O was 7-11.5% higher for $\alpha = 0.05$ and 3.5-10% higher for $\alpha = 2.5 \times 10^{-6}$. These results are in line with the literature (Lee et al., 2012).

Regarding the comparison of the SMT with MMTs, a first observation was that the power differences between the SMT and MMTs were very small for a genome-wide scan ($\alpha = 2.5 \times 10^{-6}$), and more pronounced for candidate-gene testing ($\alpha = 0.05$) when using a sample size of $n = 1,000$. The same tendencies held for both nominal significance levels, and the further most important determinants of which test has higher power were the effect size and the percentage of causal variants. As main results, the SMT had consistently the highest power when the effect sizes were higher (c=0.6), for all percentages of causal SNVs and both nominal significance levels (except when there were 50% causal SNVs in a gene and $\alpha = 2.5 \times 10^{-6}$. When the effect sizes were moderate (c=0.3) and small (c=0.2), the power of SMTs and MMTs were very similar when there were 5% or 10% causal SNVs, SKAT/SKAT-O had slightly higher power when there were 20%, and larger power when there were 50% causal SNVs. For a given effect size, increasing the percentage of causal variants in a gene led to a power increase for each test. The direction of SNV effects within a gene did not seem to greatly influence the question which test has the highest power, illustrating the robustness of SKAT/SKAT-O. For further details and sensitivity checks of the results, additional analyses, and a comparison for generated case-control traits, see sections A.4.1.1-A.4.1.3.

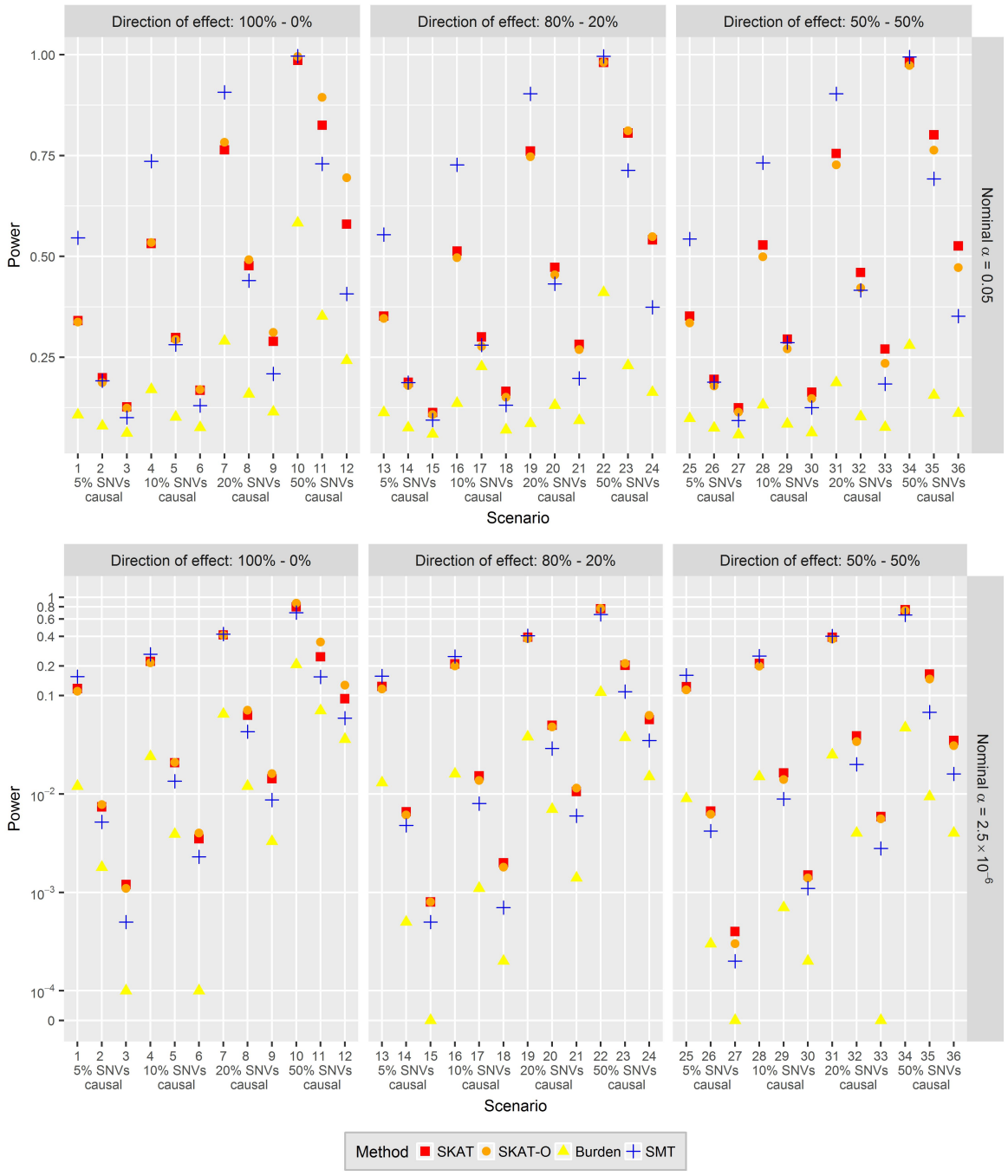


Figure 4.2: Empirical power estimates of the SMT and MMTs. Data was generated under an alternative-hypothesis model described in scenarios 1-36 in Table 4.1 of size $n = 1,000$ for $m = 10,000$ replicates. The nominal α was set to 0.05 (upper panel) and 2.5×10^{-6} (lower panel). In the lower panel with $\alpha = 2.5 \times 10^{-6}$, the coordinate system is shown on a $\log_{10}$-scale to better visualize the small power differences between the approaches. Multiple testing corrections for the SMT of all SNVs in a gene were done using the BH-correction.

4.2 C-JAMP

In this section, the validity of the parameter and standard error estimates, the test statistic, and the empirical type I error and power of hypothesis tests of genetic effects under the C-JAMP model are evaluated - for the analysis of rare variants using simulation studies. Weak ($\tau = 0.2$), moderate ($\tau = 0.5$) and strong ($\tau = 0.8$) dependences between traits were considered. Since the results in the previous section showed that SMTs can have larger or equal power compared to MMTs as long as there are not a large number of causal SNVs in a region all with small effect sizes, C-JAMP will be applied as a SMT. C-JAMP is compared to the same SMT and MMTs as described in section 4.1, to the multivariate SMTs MultiPhen (O'Reilly et al., 2012) and aSPU (Kim et al., 2016), as well as to the multivariate MMTs MURAT (Sun et al., 2016b), aSPUset and aSPUset-Score (Kim et al., 2016).

4.2.1 Material and methods

The same underlying genetic data as described in section 4.1.1.1 was used for the simulation study, with $m = 10,000$ replicates of $k = 1,000$ genotypes of $n = 1,000$ individuals. Regarding the phenotypic data, two traits Y_1 and Y_2 were generated from a bivariate distribution with weak ($\tau = 0.2$), moderate ($\tau = 0.5$), and strong ($\tau = 0.8$) dependence between the traits using the Clayton copula model in (3.1.2), conditional on a binary covariate Z_1 and a normally distributed covariate Z_2 (and conditional on the causal SNVs under the alternative hypotheses) analogous to the description in section 4.1.1.2 with the same marginal models:

$$Y_1 = 0.5z_1 + 0.5z_2 + \sum \beta_j x_j + \varepsilon; \qquad Y_2 = 0.5z_1 + 0.5z_2 + \sum \beta'_j x_j + \varepsilon' \tag{4.2.1}$$

with $Z_1 \sim \text{Bin}(0.5)$, $Z_2 \sim N(0,1)$, $\varepsilon, \varepsilon' \sim N(0,1)$, $\beta_j = c_{Y_1} \cdot |\log_{10}(\text{MAF}_j)|$, $\beta'_j = c_{Y_2} \cdot |\log_{10}(\text{MAF}_j)|$. Different values were considered for c_{Y_1} and c_{Y_2}. In addition to the scenarios 0-12 in the comparison of SMTs and MMTs in Table 4.1 with identical effect sizes c_{Y_1}, c_{Y_2}, five additional scenarios were considered with different effect sizes on the second trait (see Table 4.4).

To evaluate C-JAMP, joint models of the generated phenotypes Y_1 and Y_2 given the SNV x and covariates z_1, z_2 were fitted in the Clayton copula model in (3.1.2) with the normal marginal models in (3.1.7)-(3.1.8) using the `CJAMP` R package. In the application of C-JAMP, gene-level estimates of the empirical type I error and power are investigated, testing the hypothesis in (2.2.2) that the gene shows a signal for being associated with the trait. Type I errors were estimated under the null hypothesis in scenario 0 using $m = 100,000$ replicates. For the evaluation of the empirical power under scenarios 1-17, $m = 10,000$ replicates were analyzed. Causal variants were randomly chosen among all rare variants with MAF≤ 0.03 and all SNVs in the gene were analyzed. For scenarios 0-12, all results regarding type I error and power estimates are similar for the two traits Y_1 and Y_2, and are reported for testing the association with Y_1. Scenarios 13-17 were analyzed to investigate the power of C-JAMP on the first trait Y_1 when there is no or only a smaller genetic effect on the second trait Y_2, and the power is reported with respect to Y_1. In addition, the type I error rates for Y_2 are investigated in scenarios 13-15.

Investigation	Scenario	% of causal variants	Effect size weights	Direction of effect: Positive / Negative
Type I error	0	0%	$c_{Y_1} = c_{Y_2} = 0$	–
Power	1-12	as in Table 4.1	$c_{Y_1} = c_{Y_2}$ as in Table 4.1	100% / 0%
Additional scenarios of type I error and power	13	10%	$c_{Y_1} = 0.6; c_{Y_2} = 0$	100% / 0%
	14	20%	$c_{Y_1} = 0.3; c_{Y_2} = 0$	100% / 0%
	15	50%	$c_{Y_1} = 0.2; c_{Y_2} = 0$	100% / 0%
	16	10%	$c_{Y_1} = 0.6; c_{Y_2} = 0.1$	100% / 0%
	17	10%	$c_{Y_1} = 0.6; c_{Y_2} = 0.2$	100% / 0%

Table 4.4: Overview of the scenarios considered in the simulation study of C-JAMP, for the type I error and power comparison of C-JAMP with SMTs and MMTs. The scenarios vary the percentage of causal variants, their effect size, and the percentage of causal variants with effects in positive/ negative direction. The percentage of causal rare variants is with respect to the total number of rare variants with MAF≤ 0.03 in the gene. The effect size of a variant X_j with a given MAF on the trait Y_l is $\beta_j = c_{Y_l} \cdot |\log_{10}(\text{MAF}_j)|$.

The results of the univariate SMT, burden test, SKAT, and SKAT-O for scenarios 0-12 have already been described in section 4.1 and are reprinted below for the comparison with C-JAMP. The multivariate single- and multi-marker tests MURAT, MultiPhen, aSPU, aSPUset, and aSPUset-Score are all evaluated under the null hypothesis in scenario 0 to investigate their empirical type I errors based on $m = 10,000$ replicates. Only the aSPUset and aSPUset-Score tests provided empirical type I errors close to the nominal level and were therefore investigated in the power comparison for scenarios 1-12 on $m = 10,000$ replicates. Since the aSPU and aSPUset tests compute p-values based on simulations or permutations, they are very computationally intensive and could only be evaluated for a nominal α level of 0.05. For example, the fitting and testing of one model analyzing 24 SNVs in data of size $n = 1000$ with 10,000,000 permutations, which would be necessary to obtain p-values up to 10^{-8}, took 27 hours on a computing cluster using 100GB memory.

For more details regarding the multivariate tests, the *multivariate rare-variant association test* (MURAT; Sun et al., 2016b) is a multivariate generalization of the multi-marker test SKAT, and derives a data-adaptive variance-component test based on a Score-type test statistic, testing the overall effect of all SNVs on all traits. It is based on a multivariate version of the linear mixed model in (4.1.5) with multivariate $\boldsymbol{Y}$ including p traits, assuming that the effects $(\beta_{XY,j,l})_{j=1,\ldots,k;l=1,\ldots,p}$ of SNVs X_j on traits Y_l follow a multivariate normal distribution, with a common correlation ρ between the effects of the same SNV on the different traits, $cor(\beta_{XY,l,j}, \beta_{XY,l',j}) = \rho$ for all $1 \leq l, l' \leq p$, and uncorrelated genetic effects of different SNVs, $cor(\beta_{XY,l,j}, \beta_{XY,l,j'}) = cor(\beta_{XY,l,j}, \beta_{XY,l',j'}) = 0$ for all $j \neq j', 1 \leq l, l' \leq p$. MURAT was computed using the `MURAT()` function in the R package of the same name (available through `install_github("GreenwoodLab/MURAT")`), with default settings and the weight matrix in the test statistic based on the Beta distribution (see Supplementary Figure A.1 for an illustration), with weights $w_j \sim \text{Beta}(\text{MAF}_j, 1, 25)$ as suggested in Sun et al. (2016b).

MultiPhen (O'Reilly, 2012) inverts the standard approach of regressing phenotypes on genotypes, and proposes to use proportional odds logistic regression to predict each geno-

type X_j by the multiple phenotypes $Y_1, \ldots, Y_p$. MultiPhen was computed using the `mPhen()` function in the `MultiPhen` R package with default settings, which returns SMTs of the regression coefficients, testing the association of each SNV with each trait separately. In addition, a LRT is performed to test the association of each SNV X_j with all traits, i.e. testing $H_0 : \beta_{XY,j,1} = \cdots = \beta_{XY,j,p} = 0$ where $\beta_{XY,j,l}$ is the effect of the phenotype Y_l on X_j ('joint model' in the R package). Hence, MultiPhen performs SMTs, and the minP approach was used to obtain gene-level tests of the null hypothesis in (2.2.2), with a Bonferroni correction to adjust for the multiple tests of SNVs in a gene.

Finally, the multivariate aSPU, aSPUset, and aSPUset-Score tests are multivariate data-adaptive tests which obtain the optimal combination of different powered score test statistics. For this, they fit a multivariate marginal (generalized) linear model of the p traits $Y_1, \ldots, Y_p$ conditional on one SNV (aSPU) or multiple SNVs (aSPUset, aSPUset-Score) using GEEs. The test statistics are then derived as follows. Let $\boldsymbol{U_\beta}$ denote the subcomponents of the score vector with respect to the parameters of the genetic effects of interest. Then, the multivariate single-marker aSPU test of a SNV X_j with all p traits takes the minimum p-value over all p-values p_κ, $\kappa \in \mathcal{K}$, which are obtained based on the powered test statistics

$$SPU\left(\kappa\right) = \sum_{l=1}^{p} \left(U_{\beta,j,l}\right)^{\kappa}.$$

Hence, the aSPU test chooses adaptively the optimum from the different powered test statistics $SPU\left(\kappa\right)$. Here, SPU(1) yields a burden-type test, SPU(2) yields a variance-component-type test, and SPU(∞) is closely related to the TATES approach (van der Sluis et al., 2013). Kim et al. (2016) suggest to use $\mathcal{K} = \{1, 2, \ldots, 8, \infty\}$, and the p-values are computed through simulations or permutations. For an extension as a multivariate MMT, aSPUset takes the minimum p-value p_{κ_1,κ_2} over all $\kappa_1 \in \mathcal{K}_1, \kappa_2 \in \mathcal{K}_2$, which amounts to the most powerful powered score test over all SNVs and traits. More specifically,

$$SPU\left(\kappa_1, \kappa_2\right) = \sum_{l=1}^{p} \left[\left(\sum_{j=1}^{k} \left(U_{\beta,j,l}\right)^{\kappa_1} \right)^{1/\kappa_1} \right]^{\kappa_2}.$$

Finally, the aSPUset-Score test additionally incorporates the GEE score test statistic, and derives the p-value as the minimum of the aSPUset and GEE-Score tests. As described in Kim et al. (2016), the GEE-Score test is equivalent to the MANOVA. In the simulation study, the three tests were computed using the `GEEaSPU()` and `GEEaSPUset()` functions in the `GEEaSPU` R package (Kim & Pan, 2016) with powers $\mathcal{K}_1 = \mathcal{K}_2 = \{1, 2, \ldots, 8, \infty\}$, an unstructured correlation structure of Y_1, Y_2, and otherwise default settings. aSPUset and aSPUset-Score directly provide multivariate gene-level tests of the null hypothesis in (2.2.2), i.e. testing the absence of effects of all SNVs on all traits, and aSPU was evaluated with the minP approach using a Bonferroni correction for testing all SNVs within a gene to yield gene-level p-values.

4.2.2 Results

4.2.2.1 Evaluation of C-JAMP parameter estimates and test statistic

First, I investigated the parameter estimates and standard error estimates of the genetic effect provided by C-JAMP. Slightly biased point estimates were observed for the effect of SNVs with very few minor alleles (especially for MAC $= 1, 2$) in addition to deflated/inflated standard error estimates for Kendall's $\tau \geq 0.5$, which can be expected in such situations when parameter values on the boundary of the parameter space are of interest (Supplementary Table A.12). Accordingly, the asymptotic distribution assumption for the Wald-type test may not hold for testing such SNVs with very few alleles (Supplementary Figure A.3 and Supplementary Tables A.13-A.14), and maximum likelihood (ML) estimation as well as the large-sample Wald-type test statistic require modification in such finite-sample settings. As a consequence, adjusted Wald test statistics were obtained for Kendall's $\tau \geq 0.5$ based on an empirical study, which is described in more detail in section A.4.2.1. Supplementary Figure A.5 shows histograms of the adjusted Wald test statistics, which indicate that they are well-calibrated. Consequently, the adjusted Wald-test statistics were used for all analyses of $\tau = 0.5, 0.8$ in the simulation study that are reported in the following.

4.2.2.2 Empirical type I error rates of C-JAMP and other multivariate approaches

Next, the empirical type I error of C-JAMP based on the Wald test statistics (unadjusted Wald test statistics for $\tau = 0.2$, and adjusted Wald test statistics for $\tau = 0.5, 0.8$) were investigated for different nominal levels, to assess whether they are appropriately calibrated. The results are shown in Table 4.5 and indicate that the empirical type I errors are generally close to the nominal levels. The Bonferroni correction provides empirical levels closer to the nominal levels compared to the BH correction and is therefore used in the power evaluation in the next section.

Additional investigations of C-JAMP confirmed that the empirical type I errors are also well-calibrated when evaluated on a SNV-level instead of the gene-level (Supplementary Table A.16). All the reported type I error estimates are for testing the association with the first trait, Y_1, and the results are similar to those when investigating the association with the second trait, Y_2, when none of the two traits are associated with any rare SNV in the gene (scenario 0 described in Table 4.4). In the scenario that SNVs are only affecting one of the two traits in the joint model, the empirical type I errors are valid when there is a weak or moderate dependence between the traits. If the traits have a strong dependence, however, then the type I errors are slightly inflated (Supplementary Table A.18). Finally, the empirical type I errors were investigated when the copula function is misspecified and does not correctly model the observed dependence structure between the traits, by generating traits from the bivariate normal distribution. In the analysis, the 2-parameter copula was used and did not show inflated type I errors (Supplementary Table A.17). However, the correctness of the chosen copula function should still always be checked in a first step. See section A.4.2.2 for further details.

Nominal α	Dependence	C-JAMP Bonferroni correction	C-JAMP BH correction
5×10^{-2}	$\tau = 0.2$	5.0×10^{-2}	5.6×10^{-2}
	$\tau = 0.5$	4.6×10^{-2}	5.1×10^{-2}
	$\tau = 0.8$	4.9×10^{-2}	5.4×10^{-2}
10^{-2}	$\tau = 0.2$	1.1×10^{-2}	1.2×10^{-2}
	$\tau = 0.5$	1.0×10^{-2}	1.1×10^{-2}
	$\tau = 0.8$	1.1×10^{-2}	1.2×10^{-2}
10^{-3}	$\tau = 0.2$	1.2×10^{-3}	1.3×10^{-3}
	$\tau = 0.5$	1.2×10^{-3}	1.3×10^{-3}
	$\tau = 0.8$	1.4×10^{-3}	1.5×10^{-3}
10^{-4}	$\tau = 0.2$	1.8×10^{-4}	2.0×10^{-4}
	$\tau = 0.5$	1.6×10^{-4}	1.7×10^{-4}
	$\tau = 0.8$	1.4×10^{-4}	1.5×10^{-4}
10^{-5}	$\tau = 0.2$	3.1×10^{-5}	3.1×10^{-5}
	$\tau = 0.5$	2.0×10^{-5}	2.0×10^{-5}
	$\tau = 0.8$	1.0×10^{-5}	1.0×10^{-5}

Table 4.5: Empirical type I error estimates of C-JAMP for different nominal α levels. Data was generated from the null model described in scenario 0 in Table 4.4 for $n = 1,000$ individuals with $m = 100,000$ replicates, and under different dependencies between the two traits Y_1, Y_2 (Kendall's $\tau = 0.2, 0.5, 0.8$). Adjustments for multiple testing of all SNVs in a gene were done using the Bonferroni or the BH correction. Type I error estimates are for testing the association with the first trait, Y_1, based on the adjusted Wald test statistics described in section A.4.2.1.

Next, the empirical type I errors of the other multivariate tests were evaluated (see Table 4.6). The results showed that MURAT, the joint test of all traits in MultiPhen, and the aSPU test all lead to inflated or highly inflated type I errors. The inflated type I errors of MURAT could be slightly decreased to around 0.1-0.15 when the analysis was restricted to rare variants with moderate MAF (e.g., with MAC > 5), but was still invalid. The test of the genetic association with one trait in MultiPhen, on the other hand, provided highly deflated type I errors and is also not valid. The aSPUset-Score test showed a slight inflation and the aSPUset test was the only test with valid empirical type I errors. As a result, only the aSPUset and aSPUset-Score tests were included in the power study.

4.2.2.3 Empirical power of C-JAMP compared to other approaches

For the power evaluation in the following, the adjusted Wald test statistics were used for C-JAMP with the Bonferroni correction to account for multiple testing of SNVs within a gene. The BH-correction was used for the SMT. Figure 4.3 shows the power estimates of C-JAMP and the univariate SMT and MMTs, and Figure 4.4 shows the power of C-JAMP and the multivariate MMTs, under the scenarios 1-12 described in Table 4.4, when the nominal type I errors are 0.05 or 2.5×10^{-6} (see Supplementary Tables A.19-A.20 for more details).

Dependence	MURAT	MultiPhen Y_1	MultiPhen Joint	aSPU	aSPUset	aSPUset-Score
$\tau = 0.2$	0.290	0.019	0.298	0.100	0.051	0.054
$\tau = 0.5$	0.210	0.002	0.278	0.101	0.052	0.058
$\tau = 0.8$	0.127	0.004	0.225	0.090	0.053	0.056

Table 4.6: Empirical type I error estimates of the considered multivariate tests for a nominal α level of 0.05. Data was generated from the null model described in scenario 0 in Table 4.4 for $n = 1,000$ individuals with $m = 10,000$ replicates, and under different dependencies between the two traits Y_1, Y_2 (Kendall's $\tau = 0.2, 0.5, 0.8$). Adjustments for multiple testing of all SNVs in a gene with MultiPhen and aSPU were done using the Bonferroni correction. Type I error estimates are for testing the association with the first trait Y_1 in 'MultiPhen Y_1', and for testing the joint association with both Y_1 and Y_2 in all other tests.

The results in Figure 4.3 show that first, the power of C-JAMP increases with increasing dependence between the traits. Second, in comparison to the SMT based on a univariate model of the traits, C-JAMP leads to consistently higher power, and the power gain is larger when there is a higher dependence between the two traits. Comparing C-JAMP with the univariate MMTs when two strongly dependent traits are incorporated into the analysis, C-JAMP is always more powerful than all univariate MMTs except when 50% of all rare SNVs are causal and they have moderate or small effect sizes all in the same direction. Since the power of C-JAMP (and the SMT) is not affected when some SNVs have a positive and some a negative effect direction (scenarios 13-36 in Table 4.1), C-JAMP has higher power compared to all other approaches in these scenarios.

Regarding the power of the multivariate aSPUset and aSPUset-Score tests shown in Figure 4.4, they are inversely affected by the dependence between traits: if the dependence between traits increases, their power decreases. The power of aSPUset-Score is always higher compared to aSPUset. Furthermore, the power of C-JAMP is consistently much higher compared to both tests except when the dependence between traits is $\tau = 0.2$, there are 20% causal SNVs in a gene and effect sizes are high or there are 50% causal SNVs in a gene. Then, the power of aSPUset-Score is similar to C-JAMP or slightly higher (scenarios 11, 12). In comparison to SKAT-O, aSPUset-Score has higher power for $\tau = 0.2$ and large effect sizes, and smaller power in all other situations. In sensitivity checks, the SPU tests and aSPU were computed with other correlation structures for modeling the dependence between traits (unstructured, independence, exchangeable), with other power sets, and for analyzing only rare SNVs in a gene, and the results were very similar.

Furthermore, if the genetic effects are only affecting the first trait and are absent on the second trait (scenarios 13-17), then the power of C-JAMP is not affected and does not decrease (see Supplementary Table A.21). The power of aSPUset and aSPUset-Score, however, decreases markedly when the genetic effect is absent or smaller on the second trait. As a summary, the results of the simulation study showed that SMTs based on joint copula models outperform all other considered tests in almost all considered scenarios.

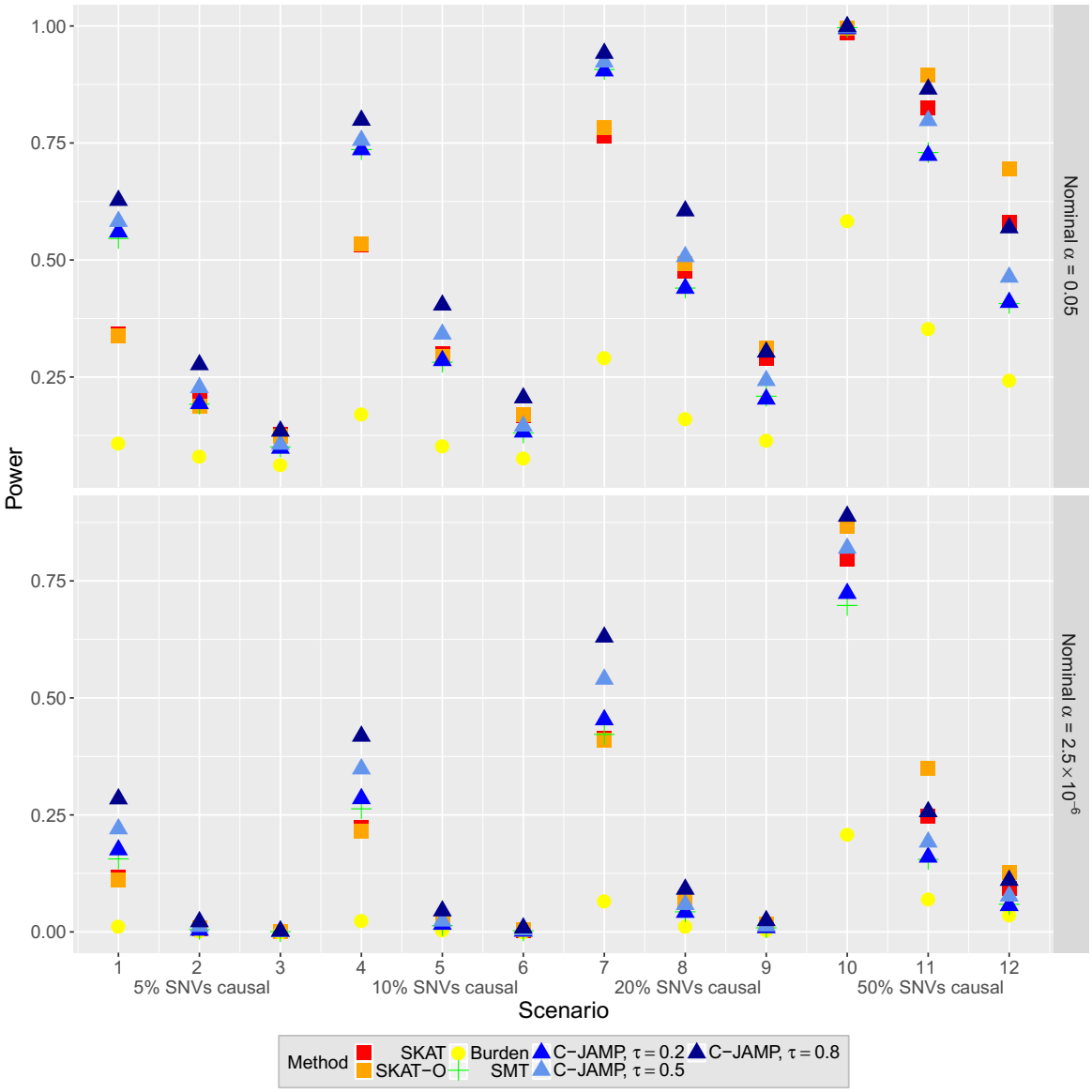


Figure 4.3: Empirical power estimates of C-JAMP versus the univariate SMT and MMTs. Data was generated under an alternative-hypothesis model described in scenarios 1-12 in Table 4.4 for $n = 1,000$ individuals with $m = 10,000$ replicates. The nominal α was set to 0.05 and 2.5×10^{-6}. Adjustments for multiple testing of all rare SNVs in a gene with SMTs were done using the Bonferroni-correction. Power estimates are for testing the association with the first trait, Y_1.

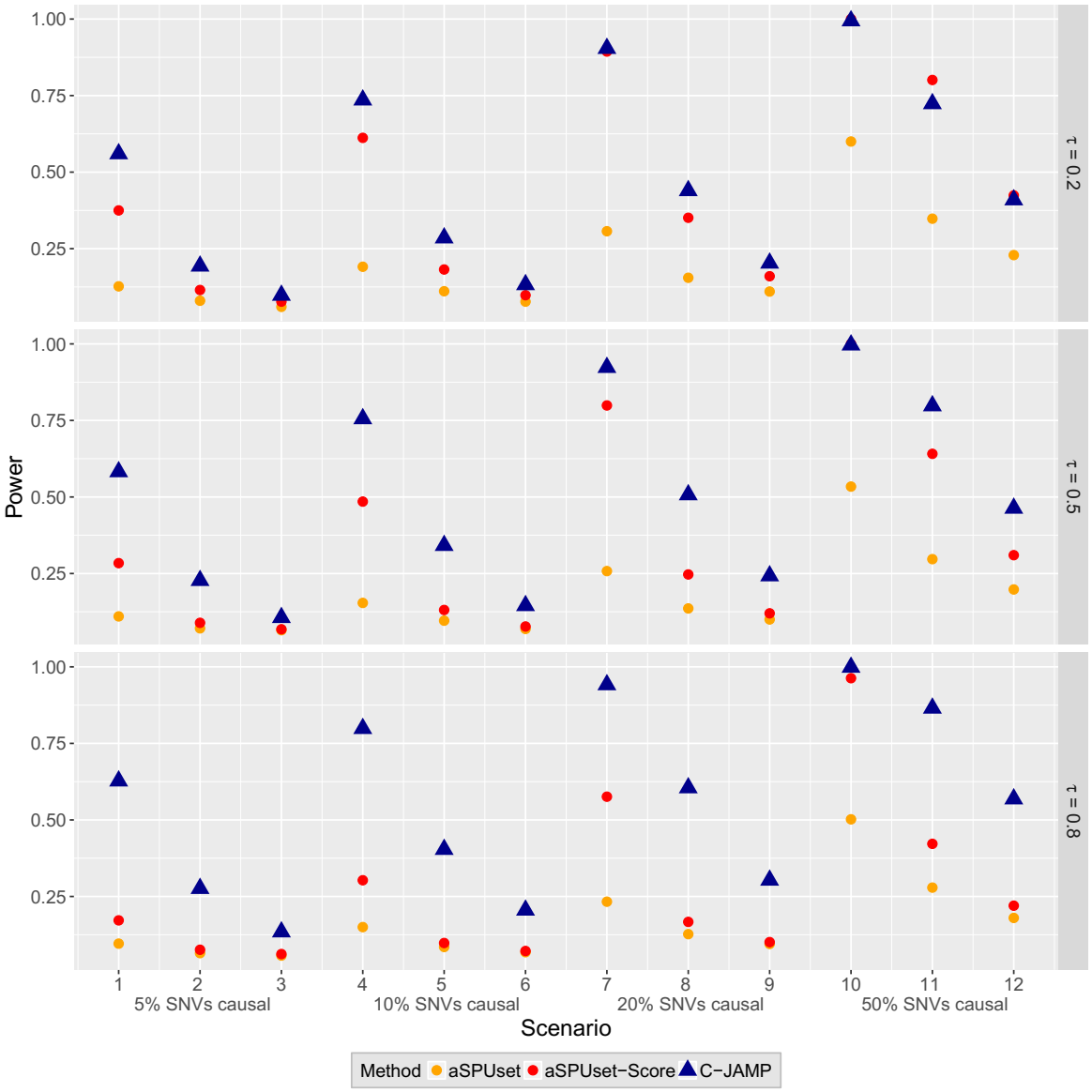


Figure 4.4: Empirical power estimates of C-JAMP versus multivariate MMTs. Data was generated under an alternative-hypothesis model described in scenarios 1-12 in Table 4.4 for $n = 1,000$ individuals with $m = 10,000$ replicates, and under different dependencies between the two traits Y_1, Y_2 (Kendall's $\tau = 0.2, 0.5, 0.8$). The nominal α was set to 0.05. Adjustments for multiple testing of all rare SNVs in a gene with C-JAMP were done using the Bonferroni-correction. Power estimates are for testing the association with the first trait, Y_1 for C-JAMP and for testing the association with both traits with aSPUset and aSPUset-Score.

4.3 CIEE

4.3.1 Material and methods

In the simulation studies to evaluate CIEE and existing methods, first, the properties of the coefficient estimates $\widehat{\alpha}_{XY}$ were investigated, and whether the effect of the intermediate phenotype K is successfully removed from the primary phenotype Y. Next, the empirical type I error and power estimates were assessed. For a quantitative primary phenotype, CIEE was compared with the two naïve regression approaches (MR and RR), the sequential G-estimation method (Vansteelandt et al., 2009) and the SEM method (Bollen, 1989; Rosseel, 2012). Under the AFT model, CIEE was compared to the naïve MR approach and the extension of the sequential G-estimation method proposed by Lipman et al. (2011). A general overview about the data generation is given in Box 4.1, and details of the scenarios and parameter values can be found in Tables 4.7-4.8 and are visualized in Figure 4.5.

For the LM setting, data was generated from

$$\begin{aligned}
U_i &\sim N\left(\mu_U,\, \sigma_U^2\right) \\
X_i &\sim \text{Bin}\left(n = 2, p = \text{MAF}_X\right) \\
L_i &= \alpha_{UL}u_i + \alpha_{XL}x_i + \varepsilon_{L,i}; \quad \varepsilon_{L,i} \sim N\left(\mu_L, \sigma_L^2\right) \\
K_i &= \alpha_{XK}x_i + \alpha_{LK}l_i + \varepsilon_{K,i}; \quad \varepsilon_{K,i} \sim N\left(\mu_K, \sigma_K^2\right) \\
Y_i &= \alpha_{UY}u_i + \alpha_{KY}k_i + \alpha_{XY}x_i + \alpha_{LY}l_i + \varepsilon_{Y,i}; \quad \varepsilon_{Y,i} \sim N\left(\mu_Y, \sigma_Y^2\right)
\end{aligned}$$

For the data generation under the AFT setting, the following additional steps were followed. First, the time-to-event phenotype $T_i = \exp(Y_i)$ was generated for each individual $i = 1, \ldots, n$. Then, the right-censoring time C_i was generated from the uniform distribution with parameters a, b so that on average, $k\%$ of the individuals were censored as pre-specified:

$$C_i \sim Unif(a, b)$$

The observed time-to-event data t_i is then the minimum of C_i and T_i, $t_i = \min(C_i, T_i)$, and the censoring indicator δ is $\delta_i = \begin{cases} 1 & \text{if } t_i = T_i \\ 0 & \text{if } t_i = C_i \end{cases}$.

The parameters a, b were chosen as described in Table 4.8.

Box 4.1: Overview of data generation in the CIEE simulation study.

The genetic marker X was generated with an additive genetic coding for minor allele frequencies $\text{MAF}_X = 0.05, 0.1, 0.2, 0.4$. The phenotypes and factors were then generated from different subgraphs of the DAG in Figure 3.4 with different effect sizes, for a sample of $n = 1,000$ individuals and using $m = 10,000$ replication datasets. Under the AFT model, time-to-event traits with 10%, 30%, and 50% censoring were considered. The effect sizes were set to simulate realistic situations with small genetic effects and small/moderate effects of the intermediate phenotype and the measured as well as unmeasured factors on the pri-

mary phenotype (scenarios 1-5). Under the LM null model, two additional scenarios were investigated where scenario 6 contains confounding of the indirect effect through measured factors, and scenario 7 equals scenario 4 but with larger effect sizes. In scenario 6, while the data generation contains a non-zero effect of L on Y, the CIEE, SEM, and sequential G-estimation methods assume $\alpha_{LY} = 0$ in the analysis, providing a test of robustness against model misspecification.

Setting	Investigation	Scenario	Censoring	MAF_X	α_{UL}	α_{XL}	α_{XK}	α_{LK}	α_{LY}	α_{UY}	α_{KY}	α_{XY}
LM	Type I error	1	–	0.05; 0.1; 0.2; 0.4	0	0	0.2	0	0	0	0.3	0
		2	–	0.05; 0.1; 0.2; 0.4	0	0.2	0.2	0.3	0	0	0.3	0
		3	–	0.05; 0.1; 0.2; 0.4	0.3	0.2	0.2	0.3	0	0	0.3	0
		4	–	0.05; 0.1; 0.2; 0.4	0.3	0.2	0.2	0.3	0	0.3	0.3	0
		5	–	0.05; 0.1; 0.2; 0.4	0.3	0.2	0.2	0.3	0	0.3	0	0
		6	–	0.05; 0.1; 0.2; 0.4	0	0	0.2	0.3	0.3	0	0.3	0
		7	–	0.05; 0.1; 0.2; 0.4	0.5	1	1	0.5	0	0.5	1	0
	Power	1	–	0.2	0	0	0.2	0	0	0	0.3	0.1; 0.2
		2	–	0.2	0	0.2	0.2	0.3	0	0	0.3	0.1; 0.2
		3	–	0.2	0.3	0.2	0.2	0.3	0	0	0.3	0.1; 0.2
		4	–	0.2	0.3	0.2	0.2	0.3	0	0.3	0.3	0.1; 0.2
		5	–	0.2	0.3	0.2	0.2	0.3	0	0.3	0	0.1; 0.2
AFT	Type I error	1	10%; 30%; 50%	0.2	0	0	0.2	0	0	0	0.3	0
		2	10%; 30%; 50%	0.2	0	0.2	0.2	0.3	0	0	0.3	0
		3	10%; 30%; 50%	0.2	0.3	0.2	0.2	0.3	0	0	0.3	0
		4	10%; 30%; 50%	0.2	0.3	0.2	0.2	0.3	0	0.3	0.3	0
		5	10%; 30%; 50%	0.2	0.3	0.2	0.2	0.3	0	0.3	0	0
	Power	1	30%	0.2	0	0	0.2	0	0	0	0.3	0.1; 0.2
		2	30%	0.2	0	0.2	0.2	0.3	0	0	0.3	0.1; 0.2
		3	30%	0.2	0.3	0.2	0.2	0.3	0	0	0.3	0.1; 0.2
		4	30%	0.2	0.3	0.2	0.2	0.3	0	0.3	0.3	0.1; 0.2
		5	30%	0.2	0.3	0.2	0.2	0.3	0	0.3	0	0.1; 0.2

Table 4.7: Overview of the scenarios considered in the simulation study of CIEE. In all scenarios, data was generated for $n = 1,000$ individuals and $m = 10,000$ replicates, with $\alpha_{UX} = 0$, $\mu_U = \mu_L = \mu_K = \mu_Y = 0$, $\sigma^2_U = \sigma^2_L = \sigma^2_K = \sigma^2_Y = 1$.

Investigation	Scenario	Censoring	α_{XY}	a	b
Type I error	1	10%; 30%; 50%	0	0.3; 0.2; 0	14.75; 4.55; 2.48
	2	10%; 30%; 50%	0	0.3; 0.2; 0	14.90; 4.58; 2.49
	3	10%; 30%; 50%	0	0.3; 0.2; 0	15.00; 4.58; 2.49
	4	10%; 30%; 50%	0	0.3; 0.2; 0	15.80; 4.70; 2.50
	5	10%; 30%; 50%	0	0.3; 0.2; 0	14.25; 4.40; 2.42
Power	1	30%	0.1; 0.2	0.2	4.75; 5.00
	2	30%	0.1; 0.2	0.2	4.81; 5.05
	3	30%	0.1; 0.2	0.2	4.80; 5.05
	4	30%	0.1; 0.2	0.2	4.94; 5.17
	5	30%	0.1; 0.2	0.2	4.60; 4.85

Table 4.8: Overview of the parameters a, b in the Uniform distribution $Unif(a, b)$ used to generate censoring times in the simulation study of CIEE.

For CIEE, sandwich standard error estimates were obtained and Wald-type test were performed. Also, standard errors were obtained using nonparametric bootstrap (with $B = 1,000$ resamples). For the two traditional approaches, MR and RR, estimates of α_{XY} were

obtained from fitting the models in equations (3.2.17)-(3.2.19), and hypothesis tests of $H_0 : \alpha_{XY} = 0$ versus $H_A : \alpha_{XY} \neq 0$ were performed by using the `mult_reg()` and `res_reg()` functions in the `CIEE` R package as described in section 3.2.2.2. In order to obtain estimates of α_{XY} and its standard error estimate under the SEM method and test $H_0 : \alpha_{XY} = 0$ versus $H_A : \alpha_{XY} \neq 0$, the `sem()` function in the `lavaan` R package (Rosseel, 2012) was used through the `sem_appl()` function in the `CIEE` R package with default settings as described in section 3.2.2.2. To apply the sequential G-estimation methods, the functions `CGcont()` and `CGsurvreg()` in the R package `CGene` (Lipman & Lange, 2011), obtained from `https://cran.r-project.org/src/contrib/Archive/CGene/`, were used with default values and adapted to the considered log-linear model for the analysis.

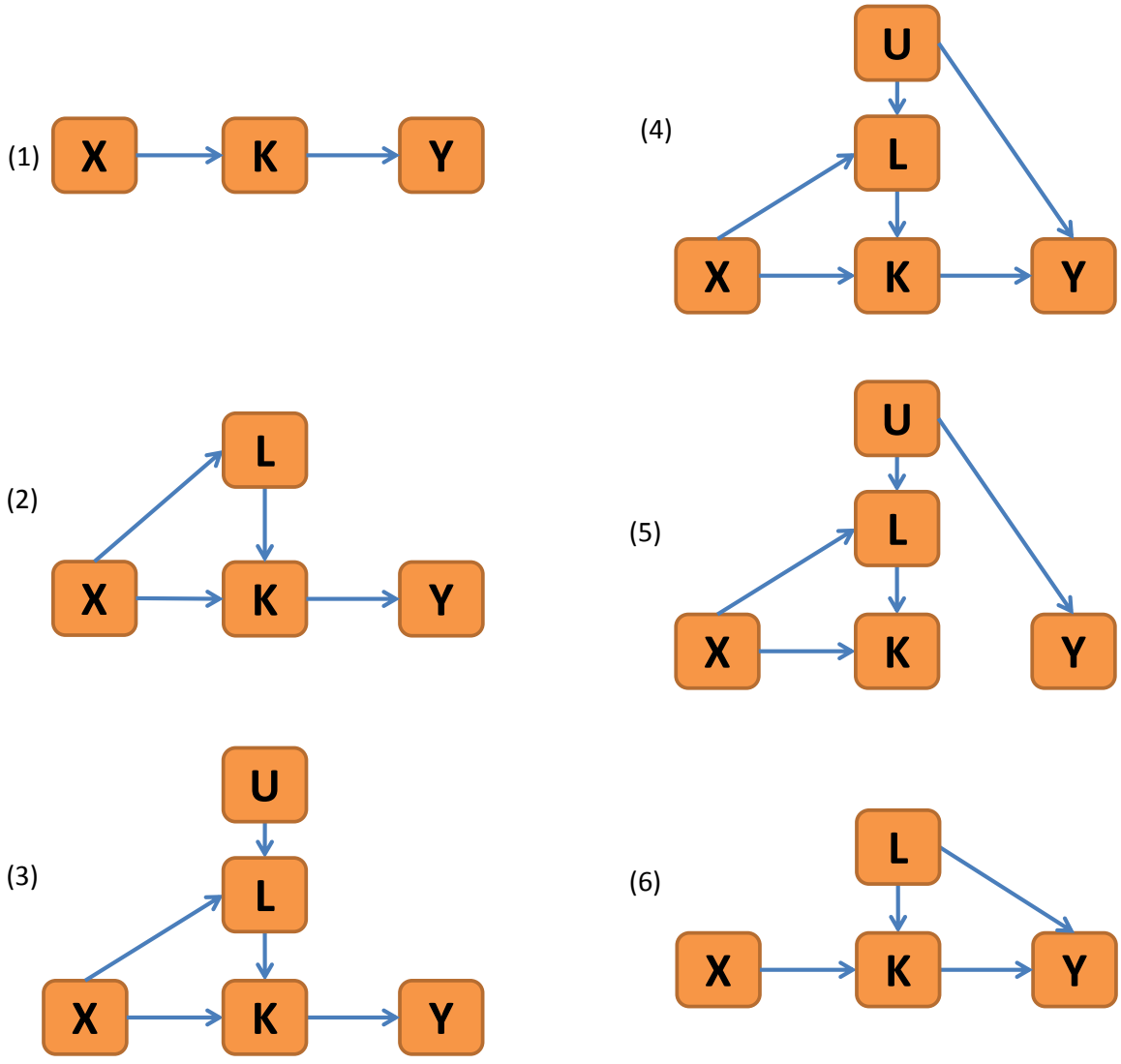


Figure 4.5: Overview of the scenarios considered in the simulation study of CIEE. The models are sub-models of the DAG in Figure 3.4 with some of the effects set to 0. Scenario 7 in the LM setting equals scenario 4 in this figure with larger effect sizes. Scenario 6 in the LM setting contains a non-zero effect of L on Y in the data generation, providing a test of robustness against model misspecification. Nonzero direct effects of X on Y are considered under each scenario to investigate the power of the approaches.

4.3.2 Results

4.3.2.1 Estimation of coefficients and standard errors

First, the coefficient estimates $\widehat{\alpha}_{XY}$ of the direct genetic effect and their standard error estimates were investigated for all methods, for the analysis of quantitative and time-to-event primary phenotypes, under the null and alternative hypotheses (see Supplementary Tables A.22-A.25). The results showed that the CIEE point estimates of the direct genetic effect are unbiased across all scenarios. Also, the standard error estimates based on the estimating equations' Huber-White sandwich estimate, nonparametric bootstrap, and the empirical standard deviation of point estimates (Supplementary Table A.26) were identical up to 2 decimals. Further checks showed that the effect of K on Y was successfully removed using the CIEE method so that $\tilde{Y}$ is uncorrelated with K (data not shown).

Regarding the naïve approaches, the coefficient estimates under the MR and RR models showed some bias whenever there was some unmeasured confounding (scenarios 4, 5 and 7 in Supplementary Table A.22 under the null hypothesis in the LM setting). The direct effect can be underestimated as in the scenarios considered here, or overestimated if, for example, the unmeasured confounding effect of U on Y is negative. When the effect of the intermediate on the primary phenotype is only confounded through measured factors in scenario 6, then both methods provide unbiased genetic effect estimates. The SEM genetic effect estimates only showed some bias when there was a higher amount of unmeasured confounding (scenario 7 under the null hypothesis in the LM setting), or when the DAG model was misspecified (scenario 6 under the null hypothesis in the LM setting, when the estimation falsely assumed $\alpha_{LY} = 0$ while the data was generated with $\alpha_{LY} = 0.3$). However, when the model was changed to correctly model an effect of L on Y in scenario 6, then unbiased genetic effect estimates were obtained (data not shown). The standard error estimates of $\widehat{\alpha}_{XY}$ obtained through MR, RR and SEM were close to the CIEE standard error estimates when the amount of unmeasured confounding was small or medium. Under scenario 7, the RR modeling approach underestimated the standard errors.

Among the investigated sequential G-estimation approaches, the method for analyzing quantitative traits (Vansteelandt et al., 2009) provided the same unbiased genetic effect estimates as CIEE, however, the approach for time-to-event traits (Lipman et al., 2011) does not remove the effect of the intermediate phenotype (see section A.4.3 in the appendix for further details) and provided strongly biased direct effect estimates whenever there was some effect of K on Y (Supplementary Tables A.24-A.25). In addition, the sequential G-estimation methods do not provide a standard error estimate of the estimated direct genetic effect.

4.3.2.2 Empirical type I error and power

As a direct consequence of the bias of genetic effect estimates discussed above, all investigated approaches except for the proposed CIEE method and the sequential G-estimation method for continuous traits led to inflated empirical type I errors in some scenarios (see Tables 4.9-4.10). Inference based on CIEE is valid for different MAFs of X, different effect sizes, with a small or moderate amount of censoring under the AFT model setting, and also if unmeasured confounding through L is present. Statistical inference remains also valid for heavy

censoring under the AFT model setting (e.g., 80% censoring) when there is no unmeasured confounding (data not shown). In addition, CIEE is robust against distributional misspecifications. For example, when Y given X, K, L, U is not normally distributed but follows a t_4, t_8, or log-normal distribution, estimates of α_{XY} remain unbiased and type I errors are valid (Supplementary Table A.26).

Scenario	MAF	CIEE	BS	G-EST	MR	RR	SEM
1	0.05	5.45%	5.40%	5.03%	5.35%	5.29%	5.04%
	0.1	5.26%	5.18%	5.05%	5.34%	5.23%	5.05%
	0.2	4.83%	4.88%	4.76%	4.94%	4.87%	5.34%
	0.4	5.16%	5.17%	5.12%	5.06%	4.80%	5.24%
2	0.05	5.17%	5.12%	4.77%	4.71%	4.67%	5.57%
	0.1	5.16%	5.13%	5.02%	5.12%	4.99%	4.75%
	0.2	5.01%	4.91%	4.87%	5.16%	4.90%	5.46%
	0.4	5.14%	5.14%	5.06%	4.91%	4.54%	4.86%
3	0.05	5.37%	5.27%	4.89%	4.89%	4.82%	5.18%
	0.1	5.17%	5.11%	4.99%	5.00%	4.87%	4.85%
	0.2	4.89%	4.90%	4.81%	5.25%	4.95%	5.19%
	0.4	5.06%	4.96%	4.97%	4.77%	4.38%	5.21%
4	0.05	5.44%	5.30%	4.98%	5.15%	5.10%	5.32%
	0.1	5.25%	5.21%	4.99%	5.27%	5.13%	4.87%
	0.2	4.81%	4.79%	4.73%	6.03%	5.68%	4.98%
	0.4	5.09%	5.14%	5.03%	5.90%	5.48%	5.43%
5	0.05	5.26%	5.11%	4.83%	4.94%	4.82%	5.42%
	0.1	5.08%	5.02%	4.91%	5.42%	5.23%	5.24%
	0.2	4.91%	4.93%	4.88%	6.01%	5.69%	5.42%
	0.4	5.12%	5.14%	5.07%	6.11%	5.75%	5.40%
6	0.05	5.14%	5.21%	4.57%	5.35%	5.29%	5.62%
	0.1	5.29%	5.27%	5.10%	5.13%	5.08%	5.83%
	0.2	5.03%	4.99%	4.83%	5.25%	5.01%	6.01%
	0.4	5.09%	5.04%	4.96%	4.94%	4.68%	6.33%
7	0.05	5.06%	4.97%	4.61%	36.14%	30.45%	21.33%
	0.1	5.05%	5.16%	4.94%	56.37%	45.31%	33.26%
	0.2	4.97%	4.93%	4.94%	73.78%	54.86%	45.47%
	0.4	5.18%	5.23%	5.17%	82.96%	59.54%	55.24%

Table 4.9: Empirical type I error estimates of CIEE and the other considered approaches under the null hypothesis in the LM setting. Data was generated for $n = 1,000$ individuals and $m = 10,000$ replicates. CIEE is the proposed method using estimating equations; BS is CIEE using nonparametric bootstrap standard errors; G-EST is the sequential G-estimation approach (Vansteelandt et al., 2009); MR is multiple regression; RR is residual regression; and SEM is structural equation modeling.

The traditional methods provide valid testing whenever there is no unmeasured confounding with RR being consistently more conservative (Table 4.9). SEM is slightly more robust to small unmeasured confounding but had inflated type I error for larger unmeasured con-

founding (scenario 7) or when the DAG model was misspecified (scenario 6). The sequential G-estimation method (Vansteelandt et al., 2009) leads to valid type I errors for all considered scenarios when quantitative traits are analyzed. For the analysis of time-to-event traits, however, the proposed G-estimation approach (Lipman et al., 2011) provides largely inflated type I errors across almost all scenarios (Table 4.10).

Scenario	Censoring	CIEE	BS	G-EST	MR
1	10%	5.29%	5.29%	22.81%	4.82%
	30%	5.24%	5.13%	24.98%	5.00%
	50%	5.29%	5.33%	20.24%	5.28%
2	10%	5.15%	5.45%	34.48%	5.28%
	30%	5.13%	5.29%	37.83%	5.15%
	50%	5.14%	5.20%	30.33%	4.74%
3	10%	5.10%	5.12%	34.54%	5.34%
	30%	4.94%	4.92%	37.25%	5.30%
	50%	4.88%	4.77%	30.66%	4.84%
4	10%	5.23%	5.19%	31.59%	6.07%
	30%	5.15%	5.15%	35.40%	6.17%
	50%	5.24%	5.14%	29.43%	5.68%
5	10%	5.15%	5.27%	4.94%	6.17%
	30%	4.98%	5.08%	4.80%	5.79%
	50%	4.93%	4.84%	4.33%	5.73%

Table 4.10: Empirical type I error estimates of CIEE and the other considered approaches under the null hypothesis in the AFT setting. Data was generated for $n = 1,000$ individuals and $m = 10,000$ replicates. The MAF of the marker X was set to 0.2. CIEE is the proposed method using estimating equations; BS is CIEE using nonparametric bootstrap standard errors; G-EST is the sequential G-estimation approach (Lipman et al., 2011); and MR is multiple log-linear censored regression.

For the power study, the same scenarios of the type I error study were considered under the LM and AFT models, with direct genetic effect sizes $\alpha_{XY} = 0.1, 0.2$. The results were highly consistent across all scenarios both for the analysis of quantitative traits (Table 4.11) and time-to-event traits (Table 4.12). They showed that all approaches have very similar power in each scenario where they have valid type I error. It is noteworthy that CIEE does not lose power compared to the traditional approaches in scenarios 1-3 where they have valid type I error. Furthermore, in the presence of unmeasured confounding in scenarios 4-5, the power of CIEE decreases only minimally while the traditional methods have inflated type I error (as well as lower power) and should not be applied.

Scenario	α_{XY}	CIEE	BS	G-EST	MR	RR	SEM
1	0.1	42.33%	42.26%	41.98%	43.13%	42.63%	42.31%
	0.2	94.62%	94.59%	94.54%	94.81%	94.68%	94.13%
2	0.1	42.52%	42.40%	42.22%	41.55%	40.53%	43.51%
	0.2	94.22%	94.09%	94.09%	94.18%	93.81%	94.15%
3	0.1	42.35%	42.30%	41.98%	42.85%	41.90%	42.32%
	0.2	94.20%	94.17%	94.06%	94.03%	93.68%	94.17%
4	0.1	39.90%	39.74%	39.53%	30.12%	29.30%	35.85%
	0.2	91.88%	91.88%	91.78%	87.92%	87.38%	90.26%
5	0.1	39.04%	38.99%	38.76%	28.79%	28.11%	35.56%
	0.2	92.48%	92.44%	92.38%	87.10%	86.66%	90.46%

Table 4.11: Empirical power estimates of CIEE and the other considered approaches under the alternative hypotheses in the LM setting. In all scenarios, data was generated for $n = 1,000$ individuals and $m = 10,000$ replicates. The MAF of the marker X was set to 0.2. CIEE is the proposed method using estimating equations; BS is CIEE using nonparametric bootstrap standard errors; G-EST is the sequential G-estimation approach (Vansteelandt et al., 2009); MR is multiple regression; RR is residual regression. Highlighted in red are those scenarios where the methods had inflated type I error so that their empirical power cannot be interpreted.

Scenario	α_{XY}	CIEE	BS	MR
1	0.1	38.30%	38.15%	38.15%
	0.2	91.14%	91.10%	91.30%
2	0.1	38.24%	38.02%	37.50%
	0.2	90.85%	90.86%	91.04%
3	0.1	38.15%	37.94%	37.54%
	0.2	90.79%	90.74%	90.57%
4	0.1	35.37%	35.10%	25.85%
	0.2	88.23%	88.14%	82.04%
5	0.1	35.46%	35.25%	25.90%
	0.2	88.73%	88.52%	83.49%

Table 4.12: Empirical power estimates of CIEE and the other considered approaches under the alternative hypotheses in the AFT setting. Data was generated for $n = 1,000$ individuals and $m = 10,000$ replicates for a genetic marker with $\text{MAF}_X = 0.2$, with 30% censoring. CIEE is the proposed method using estimating equations; BS is CIEE using nonparametric bootstrap standard errors; and MR is the multiple log-linear censored regression approach. Highlighted in red are those scenarios where MR had inflated type I error so that its empirical power cannot be interpreted.

Chapter 5

Applications of C-JAMP and CIEE

5.1 Genetic effects on blood pressure

In this section, the application of C-JAMP and CIEE to the Genetic Analysis 19 (GAW19) data is described, with the goals (i) to identify genetic variants associated with BP by testing in a joint model of BP and gene expression using C-JAMP, and (ii) to identify direct genetic effects on BP by removing indirect effects through gene expression using CIEE. Both approaches were compared to traditional regression approaches, and C-JAMP was compared to popular MMTs as well. In an additional study, the comparison of a simple SMT and MMTs in the analysis of the GAW19 data is described in the appendix in section A.4.1.4.

5.1.1 Material and methods

5.1.1.1 Sample characteristics and data description

The analyses described in this section are based on the family dataset of the GAW19, which contains information on 157 unrelated individuals from the San Antonio family studies pedigrees, including SBP and diastolic BP measurements, information regarding current use of antihypertensive medication, and nongenetic covariates (sex, age, and current tobacco smoking status) at one or more examination time points. Further information about the sample or other covariates were not available for analysis, see Blangero et al. (2016) for a detailed data description. Gene expression (GE) measurements in lymphocytes at the first time point are available for 20,634 transcripts, which were already quality-checked, filtered for detectable expression, and quantile-normalized. For $n = 113$ unrelated individuals, complete data is available for BP, GE, and the nongenetic covariates, and among them, $n = 81$ individuals have whole-genome sequence data. The genetic association analyses were conducted in this subset of $n = 81$ unrelated individuals, which included 34 men and 47 women, with a mean age of 53.1 years (SD=14.9 years) and a mean systolic blood pressure of 127.9 mmHg (SD=20.5 mmHg). 26% of the individuals were smoking, and 20% took anti-hypertensive medication. The sequence data is available for odd-numbered chromosomes and was first processed before the analysis, with standard quality control checks and the exclusion of SNVs with more than 20% missing base calls. BP measurements and nongenetic covariates are analyzed at the first time point and of the two BP phenotypes, only SBP is considered.

In the analysis, I focused on SNVs on chromosome 19, as it contains the transcript with

the highest association with SBP (IL12RB1, Kendall's τ=0.24, p=2.5×10^{-4}) among all 11,542 available mapped transcripts, and it can be hypothesized that its analysis could lead to the largest power increase under the joint copula model compared to univariate models. The gene expression of all measured 848 transcripts on chromosome 19 was investigated in the analysis. In the C-JAMP analysis, all $k = 68,727$ non-monomorphic, biallelic SNVs with MAF equal to or greater than 0.015 (i.e., with at least 3 copies of the minor allele) lying in cis within 5kb of the 848 genes were analyzed separately in copula models of SBP and the corresponding GE. The smallest gene contained 4 SNVs and the largest gene contained 920 SNVs. Of the $k = 68,727$ variants, 18,916 were rare variants with MAF between 0.015 and 0.05 (inclusive), and the remaining 49,811 variants had a MAF greater than 0.05. In the CIEE analysis, the analysis was further restricted to SNVs with MAF greater than 0.05 which yielded $k = 45,200$ SNVs after additional quality checks for the analysis. To obtain physical positions of genes and SNVs, the annotations in the provided variant call format (VCF) files, the provided mapping of gene names, and the Ensembl database with reference genome GRCh37.p13 through BioMart were used.

5.1.1.2 Statistical details of C-JAMP analysis

In a preparatory step of phenotype definition for the C-JAMP analysis, the SBP and gene expression measures were adjusted for the effect of the nongenetic covariates, including antihypertensive medication. Adjusting SBP for the effect of BP-lowering medication is important when the objective is to explain the variation in SBP (Tobin et al., 2005; Konigorski et al., 2014). Following the method described in Konigorski et al. (2014), a censored regression model was fitted conditional on the nongenetic covariates age, sex, and smoking status, with medication use as a censoring indicator δ. For treated individuals, the "true" underlying SBP was estimated from the conditional expectation of SBP, given that the observed SBP is lower than the true SBP, and also conditional on the nongenetic covariates. For further analysis, adjusted SBP phenotypes SBP_{adj} are the residuals obtained as follows: an untreated individual's adjusted SBP is the difference between observed and fitted SBP; a treated individual's adjusted SBP is the difference between estimated true and fitted SBP. Adjusted gene expression phenotypes $GE_{s,adj}, s = 1, \ldots, 848$, for the following analysis steps are residuals obtained from fitting a linear regression for each GE_s on the nongenetic covariates age, sex, smoking status, and antihypertensive medication.

In the genetic association analysis, a copula model was fitted for the joint distribution of SBP_{adj} and $GE_{s,adj}, s = 1, \ldots, 848$, conditional on each single variant x_j in and around the gene,

$$F(SBP_{adj}, GE_{s,adj}|x_j) = C_\psi(F_1(SBP_{adj}|x_j), F_2(GE_{s,adj}|x_j)), \tag{5.1.1}$$

where C_ψ is a copula function from the 2-parameter copula family in equation (3.1.4) with dependence parameter ψ, and F_1 and F_2 are the marginal distributions of SBP_{adj} and $GE_{s,adj}$, respectively, with models

$$SBP_{adj} = \alpha_0 + \alpha_{XY} x_j + \varepsilon, \tag{5.1.2}$$

$$GE_{s,adj} = \beta_0 + \beta_{XY} x_j + \varepsilon'. \tag{5.1.3}$$

To identify SNVs associated with SBP_{adj} and $GE_{s,adj}$, the null hypotheses $H_0 : \alpha_{XY} = 0$ (vs. $H_A : \alpha_{XY} \neq 0$) and $H_0 : \beta_{XY} = 0$ (vs. $H_A : \beta_{XY} \neq 0$) were tested, respectively, for all SNVs X_j and genes G_s, by using the large-sample Wald test statistics. On average, 96 SNVs were tested with each gene expression (some SNVs were tested with more than 1 gene expression because of gene overlap/proximity).

For a comparison with univariate approaches, standard linear regression models of SBP_{adj} in equation (5.1.2) and $GE_{s,adj}$ in equation (5.1.3) were fitted independently. Aside these SMTs, MMTs (a burden test, SKAT, and SKAT-O) were also considered. To have the best possible performance of the SKAT and SKAT-O as a reference, all possible kernels, p-value computation methods, and inclusions of variants (in gene only, in and around gene, rare variants, or rare and common variants) of the `SKAT()` function in R were used, and the minimum p-value of all these tests was extracted for a given gene. To assess the significance of the results corrected for multiple testing, adjusted p-values were calculated using the R function `p.adjust()` with the BH-adjustment, which controls the false discovery rate.

5.1.1.3 Statistical details of CIEE analysis

For the CIEE analysis, SBP was chosen as the primary phenotype Y and gene expression as the secondary phenotype K that could mediate the genetic effect of SNVs X_j on Y. The primary goal was to identify SNVs with a direct effect on SBP that is not (or only partially) mediated through gene expression. I assume the underlying DAG in Figure 1.1 (right panel) and that the covariates age, sex, and smoking are not related to the SNVs under investigation, but can be confounders (denoted by L_1, L_2, L_3) of the relationship between K and Y. As described in the previous section, adjusting BP for the effect of BP-lowering medication is crucial. Hence, this data analysis illustrates an application of CIEE when the primary phenotype is subject to censoring.

Some of the analyzed 45,200 SNVs were considered for their association with more than one gene expression, since they were in close proximity to more than one gene. For each of the 53,151 tested associations, CIEE was applied under the AFT model in equations (3.2.10)-(3.2.12) with measured confounders L_1, L_2, L_3. Additionally, traditional censored regression models were computed with or without taking gene expression as secondary phenotype into account:

$$MR1 : Y_i = \alpha_0 + \alpha_1 l_{1i} + \alpha_2 l_{2i} + \alpha_3 l_{3i} + \alpha_4 k_i + \alpha_{XY} x_{ij} + \varepsilon_i \tag{5.1.4}$$

$$MR2 : Y_i = \alpha_0 + \alpha_1 l_{1i} + \alpha_2 l_{2i} + \alpha_3 l_{3i} + \alpha_{XY} x_{ij} + \varepsilon_i \tag{5.1.5}$$

5.1.2 Results

5.1.2.1 C-JAMP: Genetic effects on blood pressure and gene expression

For model selection, the Akaike Information Criterion (AIC) was computed. It was observed that the copula model had a smaller AIC than a bivariate normal model conditional on any SNV; therefore, it has a better model fit. In addition, the copula model had a smaller AIC than the working independence model when there was a moderate association between SBP_{adj} and $GE_{s,adj}$ (i.e., when the estimated Kendall's τ is greater than 0.11).

Based on an α level of 0.05 and adjusted p-values, 5 SNVs in the gene CEACAM5 are identified to be significantly associated with SBP under the copula model (Table 5.1). Additionally, 1,075 SNVs in 122 different genes are identified to be significantly associated with their corresponding gene expression as local (cis) expression quantitative trait loci (eQTLs; Table 5.2). For these significant SNVs, Kendall's τ between the corresponding gene expression and SBP is estimated as 0.21 or less. In comparison, none of the 68,727 SNVs are identified as significant based on the univariate regression model of SBP (Table 5.1), and based on the univariate regression model of GE, 800 SNVs in 80 genes are significantly associated with gene expression (Table 5.2). The power gain under the copula model compared to the univariate model is shown in Figure 5.1, which shows plots of the p-values of all tested SNVs under the copula model versus the univariate models of SBP and GE. In particular, smaller p-values are much smaller under the copula model compared to the univariate models. The power gain can also be obtained when the dependence between SBP and gene expression is very low, and it is a result of (a) smaller SE estimates of the SNV effect under the copula model, while point estimates are similar and the mean difference between point estimates is 0, and (b) the different (asymptotic) null distribution of the association test statistics (normal vs. t distribution). The top SNVs under the 2 models are in the same order with respect to p-values, and are rare variants for SBP (Table 5.1), and common variants for gene expression (Table 5.2).

SNV	MAF	Location	C-JAMP			Univariate		
			$\widehat{\alpha}_{XY}$ (SE)	p-value	Adj. p-value	$\widehat{\alpha}_{XY}$ (SE)	p-value	Adj. p-value
rs10402825	0.04	CEACAM5 (protein coding)	30.77 (6.52)	2.40×10^{-6}	4.42×10^{-2}	32.21 (6.63)	6.18×10^{-6}	1.12×10^{-1}
rs7258524	0.04	CEACAM5 (upstream)	30.75 (6.56)	2.77×10^{-6}	4.42×10^{-2}	32.28 (6.67)	6.86×10^{-6}	1.71×10^{-1}

Table 5.1: SNVs associated with SBP (i.e., with the smallest p-values in testing $H_0 : \alpha_{XY} = 0$) using C-JAMP in the GAW19 analysis, and according results using linear regression. Adj. p-value is the adjusted p-value with BH-correction. The estimated Kendall's τ between SBP_{adj} and $GE_{CEACAM5,adj}$ is -0.13. Note: The markers are in very high linkage disequilibrium (LD) ($r = 0.99$). SNVs rs34155934, rs35091611, and rs10415940 in the same gene are in perfect LD with rs7258524 and have identical association.

SNV	MAF	Location	C-JAMP			Univariate		
			$\widehat{\beta}_{XY}$ (SE)	p-value	Adj. p-value	$\widehat{\beta}_{XY}$ (SE)	p-value	Adj. p-value
rs2314667	0.32	UBA52 (upstream)	0.88 (5.42×10^{-02})	6.47×10^{-60}	1.71×10^{-55}	0.88 (5.49×10^{-02})	3.21×10^{-26}	8.71×10^{-22}
rs2314664	0.33	UBA52 (downstream)	0.85 (5.52×10^{-02})	5.01×10^{-54}	7.97×10^{-50}	0.85 (5.59×10^{-02})	7.24×10^{-25}	1.18×10^{-20}
rs2314666	0.34	UBA52 (downstream)	0.85 (5.66×10^{-02})	2.88×10^{-51}	3.81×10^{-47}	0.85 (5.73×10^{-02})	3.41×10^{-24}	4.62×10^{-20}

Table 5.2: Top 3 SNVs associated with gene expression (i.e., with the three smallest p-values in testing $H_0 : \beta_{XY} = 0$) using C-JAMP in the GAW19 analysis, and according results using linear regression. Adj. p-value is the adjusted p-value with BH-correction. The estimated Kendall's τ between SBP_{adj} and $GE_{UBA52,adj}$ is -0.01. Note: The markers in the table are in very high linkage disequilibrium (LD) (with $0.96 \leq r \leq 0.99$). SNVs rs6554 and rs10425018 in the same gene are in perfect LD with rs2314667 and have identical association; also rs7258480 is in perfect LD with rs2314664.

After extracting the smallest p-values among all SKAT and SKAT-O options in the `SKAT()` function in R and applying a multiple testing correction for p-values, 1 gene is identified to be associated with SBP (which contains the 5 significant variants identified under the copula model) and 36 genes are associated with GE. All of these genes are also identified in the

analysis with C-JAMP. Table 5.3 shows the raw and adjusted p-values of the top genes identified by SKAT and SKAT-O, and the minimum adjusted p-values of all variants in or around the corresponding gene under the copula model for comparison. Much smaller p-values are obtained using the single-variant analysis under the copula model.

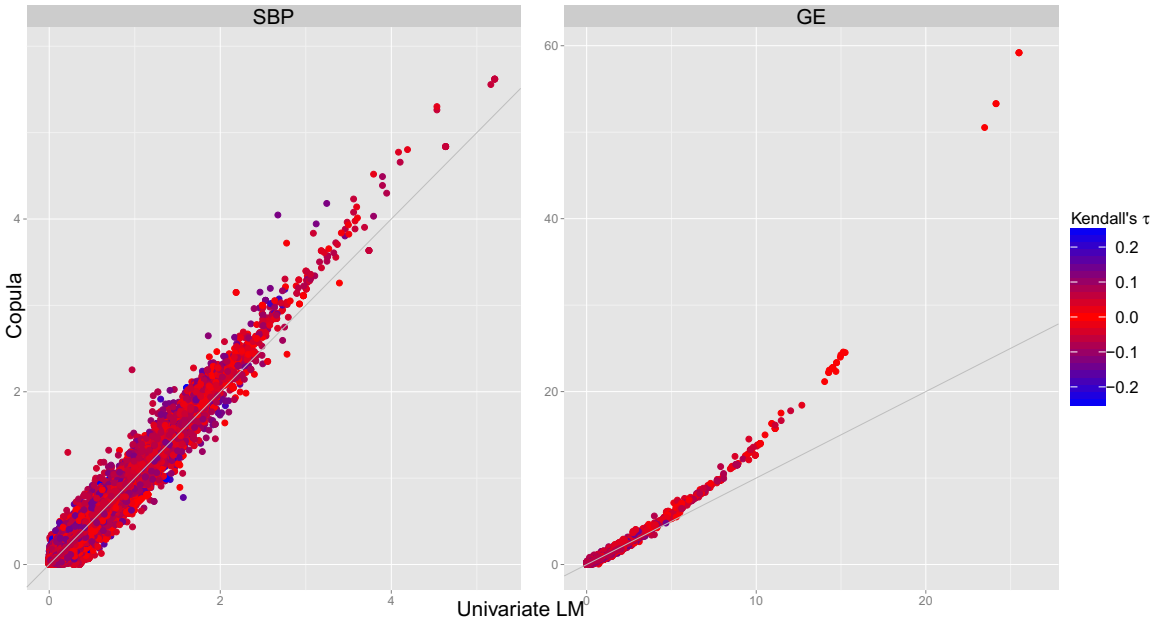


Figure 5.1: Scatterplots of p-values from C-JAMP versus linear regression from the GAW19 analysis. The p-values are from the testing of all common and rare SNVs for $H_0 : \alpha_{XY} = 0$ (for SBP, left panel) and $H_0 : \beta_{XY} = 0$ (for GE, right panel), and are shown on a $-\log_{10}$ scale. For each SNV, the strength of the association (Kendall's τ) between the gene expression (of the gene containing the SNV) and SBP is shown by the color of the dots.

The type I error of the test statistics of the 3 modeling approaches was checked using the observed data (see Figure 5.2). The p-values obtained using C-JAMP and using the univariate regression models do not appear to show inflated type I errors, and the points corresponding to high p-values lie on the diagonal line. There is some evidence for an inflated type I error for the gene-based SKAT and SKAT-O tests, when the minimum p-value of all options in the `SKAT()` function is used.

Gene	SKAT/SKAT-O		C-JAMP	
	p-value	Adjusted p-value	p-value	Adjusted p-value
UBA52	1.36×10^{-12}	1.15×10^{-09}	6.47×10^{-60}	1.71×10^{-55}
IGFLR1	3.79×10^{-10}	1.61×10^{-07}	3.74×10^{-19}	1.24×10^{-15}
ACP5	3.12×10^{-09}	8.83×10^{-07}	4.95×10^{-12}	5.33×10^{-09}
CNN2	4.96×10^{-07}	9.06×10^{-05}	2.01×10^{-14}	3.56×10^{-11}
ANKRD27	5.34×10^{-07}	9.06×10^{-05}	6.08×10^{-13}	7.94×10^{-10}

Table 5.3: Top 5 genes associated with gene expression (i.e., with the five smallest minimum p-values in testing $H_0 : \beta_{XY} = 0$ for all SNVs in/around the gene) using C-JAMP in the GAW19 analysis, with according p-values using SKAT, SKAT.

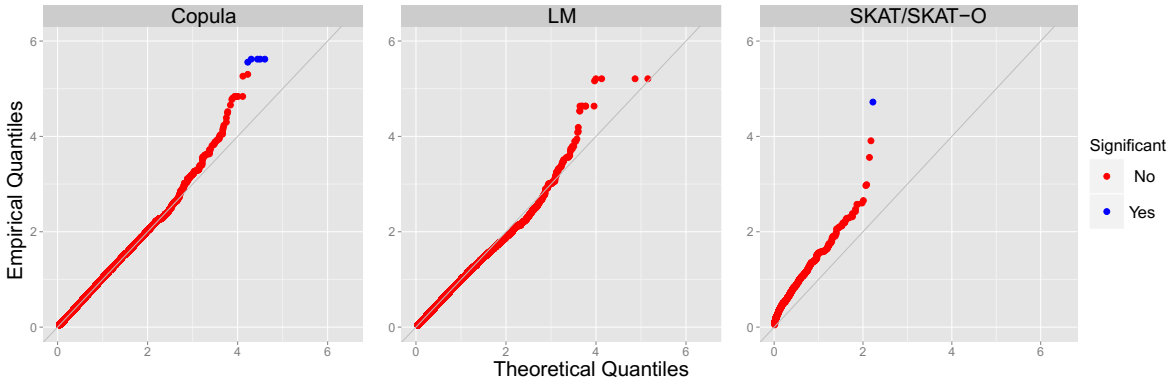


Figure 5.2: Uniform Q-Q plots of the p-values from C-JAMP, univariate regression and SKAT/SKAT-O from the GAW19 analysis. The p-values are on a $-\log_{10}$ scale from the testing of all common and rare SNVs for $H_0 : \alpha_{XY} = 0$ under the copula model (left panel), univariate model (middle panel), and SKAT/SKAT-O (right panel). SNVs/genes which are significant after correcting for multiple testing are highlighted in blue color.

5.1.2.2 CIEE: Direct genetic effects on blood pressure

Results from CIEE, MR1 and MR2 are shown for the 5 SNVs with the smallest p-values obtained from testing the absence of the direct effect on SBP using CIEE (Table 5.4). None of these SNVs were found to be associated with sex, age, or smoking (data not shown). The SNV rs56202530 with the smallest p-value is upstream of the IL27RA gene, and its direct effect on SBP is estimated to be -0.148 (SE=0.030, p-value $= 7.2 \times 10^{-7}$) using CIEE, and -0.080 (SE=0.031, p-value $= 9.5 \times 10^{-3}$) using MR1. This was the only SNV with an adjusted p-value less than 0.05 using CIEE. The results obtained through MR1 and MR2 were very similar to each other. The 5 SNVs with the smallest p-values using MR1 are shown in Table 5.5. None of these SNVs returned an adjusted p-value less than 0.05.

In a comparison of the results using traditional multiple regression and CIEE, for the SNVs in Table 5.4, the estimated direct effects were in the same direction but larger using CIEE while estimated standard errors were similar - leading to different conclusions on the statistical significance of the effect estimates. Assuming the correctness of the underlying DAG in Figure 1.1 (right panel) and using the results from the simulation study, the most plausible explanation of the effect estimate differences is that there is unmeasured confounding of the indirect effect $X \to K \to Y$ through L in opposite effect direction (e.g., $X \to Y$ negative, $U \to L$ negative, $L \to K, K \to Y, U \to Y$ positive effects). This suggests that using traditional approaches without accounting for indirect effects of secondary phenotypes and confounders might miss true causal SNVs (the SNVs 1, 2, 3 and 5 in Table 5.4).

SNV	MAF		$\widehat{\alpha}_{XY}$			$\widehat{SE}(\widehat{\alpha}_{XY})$			P-value			Adjusted p-value		
			CIEE	MR1	MR2	CIEE	MR1	MR2	CIEE	MR1	MR2	CIEE	MR1	MR2
rs56202530	0.14	IL27RA	-0.148	-0.080	-0.087	0.030	0.031	0.034	7.2×10^{-7}	9.5×10^{-3}	1.1×10^{-2}	0.038	1	1
rs3746061	0.05	BTBD2	-0.106	-0.057	-0.055	0.023	0.044	0.044	5.3×10^{-6}	2.0×10^{-1}	2.2×10^{-1}	0.281	1	1
rs60458566	0.13	AP2A1	-0.123	-0.082	-0.079	0.028	0.030	0.030	8.7×10^{-6}	5.9×10^{-3}	8.2×10^{-3}	0.461	1	1
rs62117661	0.09	KLK12	0.256	0.182	0.178	0.058	0.041	0.041	1.0×10^{-5}	8.7×10^{-6}	1.4×10^{-5}	0.552	0.461	1
rs883394	0.25	ACTN4	-0.102	-0.080	-0.062	0.023	0.022	0.023	1.1×10^{-5}	3.3×10^{-4}	6.9×10^{-3}	0.624	1	1

Table 5.4: Top 5 SNVs associated with SBP (i.e., with the five smallest p-values in testing $H_0 : \alpha_{XY} = 0$) using CIEE in the GAW19 genetic association analysis of 113,890 SNVs on chromosome 19, with the shown gene (expression) as intermediate phenotype. The SNV is described by its rs identification number. For these SNVs, point estimates, standard error estimates, raw p-values and Bonferroni-corrected (adjusted) p-values obtained through CIEE and the multiple regression approaches MR1 and MR2 are shown.

SNV	MAF		$\widehat{\alpha}_{XY}$			$\widehat{SE}(\widehat{\alpha}_{XY})$			P-value			Adjusted p-value		
			CIEE	MR1	MR2	CIEE	MR1	MR2	CIEE	MR1	MR2	CIEE	MR1	MR2
rs62117661	0.09	KLK12	0.256	0.182	0.178	0.058	0.041	0.041	1.0×10^{-5}	8.7×10^{-6}	1.4×10^{-5}	0.552	0.461	1
rs1972785	0.31	ZSCAN5A	-0.086	-0.093	-0.092	0.027	0.022	0.022	1.3×10^{-3}	1.8×10^{-5}	2.6×10^{-5}	1	0.936	1
rs62117661	0.09	KLK11	0.252	0.175	0.178	0.059	0.041	0.041	2.2×10^{-5}	1.8×10^{-5}	1.4×10^{-5}	1	0.940	1
rs10415616	0.35	ZSCAN5A	-0.071	-0.089	-0.091	0.026	0.021	0.021	5.6×10^{-3}	1.8×10^{-5}	1.3×10^{-5}	1	0.943	1
rs7248173	0.31	ZSCAN5A	-0.086	-0.093	-0.092	0.027	0.022	0.022	1.3×10^{-3}	2.2×10^{-5}	3.3×10^{-5}	1	1	1

Table 5.5: Top 5 SNVs associated with SBP (i.e., with the five smallest p-values in testing $H_0 : \alpha_{XY} = 0$) using the multiple regression model MR1 in the GAW19 genetic association analysis of 113,890 SNVs on chromosome 19, with the shown gene (expression) as intermediate phenotype. The SNV is described by its rs identification number. For these SNVs, point estimates, standard error estimates, raw p-values and Bonferroni-corrected (adjusted) p-values obtained through CIEE and the multiple regression approaches MR1 and MR2 are shown.

5.2 Genetic and transcriptomic effects on obesity traits

This section describes a study which was conducted to identify novel genetic and transcriptomic loci associated with obesity traits, by measuring SNVs in coding regions, gene expression in SAT, obesity traits through MRI, and by applying C-JAMP and CIEE. In addition, it will be shown in the analysis that using C-JAMP allows identifying more associated genetic markers and genes compared to univariate analyses, and applying CIEE to infer direct genetic effects identifies SNVs that would be missed by using traditional regression approaches.

5.2.1 Material and methods

5.2.1.1 Sample characteristics and study overview

The study was conducted within the large European Prospective Investigation into Cancer and Nutrition (EPIC) Potsdam study (Riboli & Kaaks, 1997; Boeing et al., 1999). EPIC Potsdam is an ongoing cohort study among 27,548 persons aged 35-65 at recruitment between 1994 and 1998 from the general population of the city of Potsdam and surrounding area in Germany, conducted by the German Institute of Human Nutrition (DIfE) in Potsdam-Rehbrücke, Germany. From 2010 to 2013, a randomly selected sample of 1472 participants, stratified according to gender and age groups (35-45, 45-55, 55-65 years at baseline), were

reinvited to the study center (Wientzek et al., 2014). 815 participants (≈ 3% of the original cohort) attended this re-examination, forming the so-called EPIC Potsdam substudy, and completed a detailed assessment of body composition, physical activity, diet, biofluid sampling (Gottschald et al., 2016). MRI data was available on 594 participants (Neamat-Allah et al., 2014), see section A.4.4.1 for more details, quality control checks, and descriptive statistics. The MRI scans were performed in the Department of Diagnostic and Interventional Radiology, Clinic Ernst-von-Bergmann, Potsdam, Germany, and the bioinformatic processing of images was done at the Division of Medical and Biological Informatics, German Cancer Research Center (DKFZ), Heidelberg, Germany. SAT biopsies were taken from 278 participants of in total 673 eligible participants at the DIfE study center in Potsdam, with sufficient material extracted from 200 participants (see section A.4.4.2). These 200 participants constitute the sample for this study, with 160 of them having MRI measurements. The study was approved by the Ethics Committee of the medical association of the state of Brandenburg (Germany) and all participants provided written informed consent.

From the SAT biopsies, the total RNA was extracted and its quality verified at the Molecular Epidemiology Research Group at the Max-Delbrück-Center for Molecular Medicine in the Helmholtz Association (MDC), Berlin, Germany (see section A.4.4.3). The RNA was subsequently purified and prepared for sequencing, using polyA-selection to enrich mRNA, fragmentation, and random hexamer priming for cDNA synthesis. These steps were performed at the Genetics and Genomics of Cardiovascular Diseases Research Group at the MDC Berlin. Next, the cDNA was appended with adapters for multiplexing and applied on the flow cells for sequencing on the Illumina Hi-Seq 2000 at the Genomics Platform at the MDC Berlin (see section A.4.4.4). 198 probes were sequenced with 6 samples per lane, yielding on average 64,095,856 raw reads (SD=7,518,970), with minimum 43,373,110 and maximum 85,591,020 raw reads per probe. 2 samples were sequenced in higher depth on one lane each, for an evaluation of the sequencing depth. In addition, 6 probes were re-sequenced due to minor technical problems during their run, and the repeated runs can be used for reliability analyses. After the sequencing, in a first bioinformatics step, the raw reads were demultiplexed and quality-controlled (see sections A.4.4.5, A.4.4.6). Next, the reads were aligned to hg38 (GRCh38) using `TopHat2` (Kim et al., 2013) and `Bowtie 2` (Langmead & Salzberg, 2012) and again quality-controlled (sections A.4.4.7, A.4.4.8). The aforementioned bioinformatic steps were done in collaboration with the Genetics and Genomics of Cardiovascular Diseases Research Group at the MDC Berlin. All quality controls as well as the bioinformatic steps described in the following were conducted by me. After the alignment, ≈50 million single reads (i.e., about 25 million paired reads) were available at high quality per probe, and about 330 million single reads of the two deeply sequenced probes.

In order to obtain gene expression measures, the aligned reads (after filtering low-quality reads as well as reads with multiple alignments, and only using properly aligned pairs) were counted using `htseq-count` (Anders et al. 2015) and TMM-normalized TPM counts were obtained (see sections A.4.4.9, A.4.4.10). The transcripts per million (TPM) within-sample normalization (Li & Dewey, 2011) corrects for biases due to gene length and library size and the between-sample trimmed mean of M-values (TMM) normalization (Robinson & Oshlack, 2010) computes the trimmed mean of M-values between each pair of samples to account

for potential sequencing biases. Finally, the normalized read counts were quality-controlled, low-expressed genes (expressed in less than 25% or the probands) were filtered, and the Yeo-Johnson transformation (Yeo & Johnson, 2000) was used to remove skewness and yield normally-distributed gene expression measures (section A.4.4.11). This yielded 30,917 genes for the main analysis. For all analyses described in the following, the Yeo-Johnson transformed TMM-normalized TPM counts were used as measure of gene expression.

Finally, in order to investigate genetic variants, SNVs (in coding regions) were called from the RNA-Seq .bam files using the `samtools mpileup` tool (Li et al., 2009; Li, 2011). In total, 2,029,767 biallelic SNVs could be called as well as 79,618 multi-allelic (>2) SNVs and INDELS. After quality-checks and filtering of non-monomorphic non-autosomal SNVs, 509,009 biallelic SNVs were retained for the main analysis (section A.4.4.12). Across all SNVs and samples, the average coverage per position per sample was 6.6. Of the 509,009 SNVs, there are 122,042 singletons with a MAF of 0.0025 and 34,452 doubletons with MAF=0.005. Furthermore, 46,530 SNVs had MAF between 0.005 and 0.01, 151,114 SNVs had MAF between 0.01 and 0.05, and 153,348 SNVs had MAF greater than 0.05. In the analysis described in the following, complete data analysis was performed yielding 482,507 SNVs.

In order to reduce the phenotype dimensionality, results of a preliminary study can be used to choose those measures of body fat mass and body fat distribution as obesity traits, which might best capture the metabolic activity of AT. In the preliminary study (see sections A.4.4.13, A.4.4.14), gene expression in SAT and plasma concentrations were assessed of 6 adipokines with an established role in obesity and chronic disease, using quantitative polymerase chain reaction (PCR) and enzyme-linked immunosorbent assays (ELISA). As primary aim, adipokine plasma concentrations were predicted by AT measures and gene expression, and as secondary aim, the association of the AT measures with gene expression and plasma levels was investigated (see section A.4.5). The relevant results for this thesis showed that among the different compartments, SAT and TAT mass had the highest correlation with most adipokines. Hence, SAT mass is chosen as first obesity trait for the analysis here. Regarding body fat distribution, the ratio of SAT mass over TAT mass ($\frac{\text{SAT}}{\text{TAT}}$) is used as second obesity trait in the analysis. SAT mass was log-transformed for all further analyses so that it is normally distributed. The association of log(SAT) with $\frac{\text{SAT}}{\text{TAT}}$ was $\tau = 0.36$.

5.2.1.2 Statistical details of C-JAMP analyses

In the first part of the analysis, the goal was to identify novel genetic and transcriptomic obesity loci, by improving the power of association studies through a joint analysis of multiple obesity traits with C-JAMP. The tested models are shown in Figure 1.2 (left panel), and described in more detail in the following:

For the genetic association analysis, the following 2-parameter copula models were fitted:

$$F(Y_1, Y_2|\boldsymbol{z}) = C_\psi(F_1(Y_1|\boldsymbol{z}), F_2(Y_2|\boldsymbol{z})) \tag{5.2.1}$$

with marginal models

$$Y_1 = \alpha_0 + \alpha_1 z_1 + \alpha_2 z_2 + \alpha_3 z_3 + \alpha_4 z_4 + \alpha_5 z_5 + \alpha_{XY} x_j + \varepsilon, \tag{5.2.2}$$

$$Y_2 = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3 + \beta_4 z_4 + \beta_5 z_5 + \beta_{XY} x_j + \varepsilon', \tag{5.2.3}$$

where $Y_1 = \log(SAT)$, $Y_2 = \frac{\text{SAT}}{\text{TAT}}$, $\boldsymbol{z} = (z_1, z_2, z_3, z_4, z_5, x_j)^T$ including the SNVs X_j and covariates Z_1=sex, Z_2=age, Z_3=smoking, Z_4=physical activity, Z_5=education, and F_1 and F_2 are the marginal distributions of Y_1 and Y_2. These models were fitted sequentially with C-JAMP for all SNVs X_j, $j = 1, \ldots, 482,507$, and the large-sample Wald test statistics were computed to test the null hypotheses $H_{0,j} : \alpha_{XY} = 0$ (vs. $H_{A,j} : \alpha_{XY} \neq 0$) and $H_{0,j} : \beta_{XY} = 0$ (vs. $H_{A,j} : \beta_{XY} \neq 0$). Since the dependence between Y_1, Y_2 is $\tau = 0.36$, the Wald-type test statistics were not adjusted based on the results in section 4.2.

For the transcriptomic association study, the 2-parameter copula models in equations (5.2.1)-(5.2.3) were fitted analogously with gene expression (i.e., the Yeo-Johnson transformed TMM-normalized TPM counts) as predictors G_s:

$$F(Y_1, Y_2|\boldsymbol{z}) = C_\psi(F_1(Y_1|\boldsymbol{z}), F_2(Y_2|\boldsymbol{z})) \tag{5.2.4}$$

with marginal models

$$Y_1 = \alpha_0 + \alpha_1 z_1 + \alpha_2 z_2 + \alpha_3 z_3 + \alpha_4 z_4 + \alpha_5 z_5 + \alpha_{GY} g_s + \varepsilon, \tag{5.2.5}$$

$$Y_2 = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3 + \beta_4 z_4 + \beta_5 z_5 + \beta_{GY} g_s + \varepsilon', \tag{5.2.6}$$

and $Y_1 = \log(SAT)$, $Y_2 = \frac{\text{SAT}}{\text{TAT}}$, $\boldsymbol{z} = (z_1, z_2, z_3, z_4, z_5, g_s)^T$ including the Yeo-Johnson transformed TMM-normalized TPM counts G_s and covariates Z_1=sex, Z_2=age, Z_3=smoking, Z_4=physical activity, Z_5=education. The model in equations (5.2.4)-(5.2.6) was fitted separately for all genes G_s, $s = 1, \ldots, 30,917$, using C-JAMP. MLEs and Wald test statistics were obtained to test $H_{0,s} : \alpha_{GY} = 0$ (vs. $H_{A,s} : \alpha_{GY} \neq 0$) and $H_{0,s} : \beta_{GY} = 0$ (vs. $H_{A,s} : \beta_{GY} \neq 0$).

For comparison with univariate approaches, linear regression models of equations (5.2.2)-(5.2.3) and (5.2.5)-(5.2.6) were computed using the `lm()` function in R with default settings. Here, the gene expression analysis allows evaluating whether C-JAMP can also increase the power when testing quantitative predictors.

For an interpretation of the results, the SNVs with smallest p-values from the genetic association analysis are annotated and discussed regarding their potential function and relevance for obesity. Regarding the transcriptomic study, a follow-up analysis was conducted to further investigate the associated obesity genes. First, a gene ontology (GO; Gene Ontology Consortium, 2001) enrichment analysis (Alexa et al., 2006; Grossmann et al., 2007) was performed in order to identify which gene ontology terms are enriched in the obesity-associated genes compared to all 30,917 analyzed genes. For this, the `topGO` R package (Alexa & Rahnenführer, 2016) was used with the biological processes ("BP") subontology, pruning GO terms with less than 10 annotated genes before the enrichment analysis, and computing gene-GO term mappings based on the ensembl gene identifiers. As enrichment tests, classic Fisher's exact test and the classic as well as adapted "elim" Kolmogorov-Smirnoff (KS) gene-set enrichment analysis tests were computed (Alexa et al., 2006; Ackermann & Strimmer, 2009).

5.2.1.3 Statistical details of CIEE analysis

As will be described in the following results section, the C-JAMP analysis identified 607 autosomal genes whose SAT gene expression levels are associated with either body fat mass

(log(SAT)), body fat distribution ($\frac{\text{SAT}}{\text{TAT}}$), or both. Regarding genetic predictors, while C-JAMP identified some candidate SNVs with suggestive significance, none of the SNVs were associated with log(SAT) or $\frac{\text{SAT}}{\text{TAT}}$ with genome-wide significance. However, it might be the case that some of the investigated SNVs do have an effect on obesity, but their "indirect" effect on obesity through gene expression is in opposite direction of the "direct" effect on obesity (which includes all non-measured pathways of the SNV effect on obesity other than through gene expression). Such SNVs might not be detectable in joint models through C-JAMP or any other model evaluating overall genetic effects. Therefore, CIEE was used to follow-up the C-JAMP analysis to search for the presence of such markers.

More specifically, CIEE was used to test direct genetic effects of all 15,895 SNVs within the 607 autosomal genes identified by C-JAMP, while blocking the indirect genetic effect on obesity through gene expression and adjusting for the 5 covariates age, sex, smoking, physical activity, education (see Figure 1.2, right panel). CIEE was computed under the LM setting in equations (3.2.1)-(3.2.3) with SNVs X_j, primary phenotype Y being log(SAT) or $\frac{\text{SAT}}{\text{TAT}}$, intermediate phenotype K being the gene expression (the Yeo-Johnson transformed TMM-normalized TPM counts) of gene G_s containing SNV X_j, and the covariates $L_1, \ldots, L_5$ being age, sex, smoking, physical activity and education. For comparison, the two traditional regression approaches MR and RR were computed using the `lm()` R function with default settings, with model (5.2.7) for MR and the according model of RR as described in equations (3.2.18)-(3.2.19).

$$Y = \alpha_0 + \alpha_1 l_1 + \alpha_2 l_2 + \alpha_3 l_3 + \alpha_4 l_4 + \alpha_5 l_5 + \alpha_6 g_s + \alpha_{XY} x_j + \varepsilon. \tag{5.2.7}$$

CIEE, MR, and RR models were computed for the primary traits $Y = \log(\text{SAT})$ or $\frac{\text{SAT}}{\text{TAT}}$, for the genes $G_s, s = 1, \ldots, 607$, and the contained SNVs X_j, $j = 1, \ldots, 15,895$.

5.2.2 Results

5.2.2.1 Descriptive statistics

Descriptive statistics of the study participants regarding their personal characteristics, disease prevalence and AT measures are shown in Table 5.6. The sample contained slightly more women than men (57% women, 43% men), with an average age of 65.1 years (SD=9.0 years), a mean BMI of 27.9 (SD=4.2), and a low prevalence of cardiovascular and cardiometabolic diseases. SAT was on average 20.1 kg (SD=5.2kg) for women, and 14.8 kg (SD=4.3kg) for men.

As a validity check of the obtained gene expression measures that are analyzed in the following, correlations between RNA-Seq and pPCR gene expression estimates were computed for the 6 candidate genes that were investigated in the preliminary study described in section A.4.5. The (Pearson) correlations were $r = 0.36$ for FABP4, $r = 0.56$ for leptin receptor, $r = 0.58$ for adiponectin, and $r = 0.85 - 0.87$ for leptin, interleukin-6, and resistin, in line with previous reports in the literature (Marioni et al., 2008; Wang et al., 2009, Lee et al., 2011). As further checks, it was observed that the estimated correlations between gene expression and obesity traits were similar when using RNA-Seq or PCR. For example, the correlation between leptin gene expression and SAT mass was $r = 0.59$ using RNA-Seq and $r = 0.56$ using PCR. See section A.4.4.10 for further validity checks and details.

Measures	Women	Men
Sample size	113	87
Age, years	62.9 (9.0)	67.9 (8.2)
Smoking, %		
never	55.7	27.6
former	31.0	58.6
current	13.3	13.7
CPAI, %		
inactive	9.7	8.0
moderately inactive	27.4	31.0
moderately active	32.3	28.7
active	30.1	32.2
Occupational training, %		
no vocational training/ vocational training	41.6	36.8
technical college	27.4	11.5
university	30.1	51.7
Myocardial infarction, %	1	1
Stroke, %	0	0
Heart failure, %	0	0
Diabetes, %	1.8	11.5
Height, cm	162.2 (5.6)	174.2 (6.2)
Weight, kg	73.2 (12.6)	84.5 (11.7)
BMI, kg/m^2	27.8 (4.5)	27.9 (3.8)
WC, cm	91.9 (11.7)	102.0 (10.0)
HC, cm	106.0 (10.1)	102.2 (6.5)
WHR	0.9 (0.1)	1.0 (0.1)
MRI AT measures		
VAT, kg*	2.9 (1.2)	5.2 (1.7)
CAT, kg*	0.3 (0.1)	0.4 (0.2)
SAT, kg*	20.1 (5.2)	14.8 (4.3)
TAT, kg*	23.7 (5.4)	21.0 (5.9)

Table 5.6: Gender-stratified characteristics of the obesity study population. Values are relative frequencies, mean and SD, or *median and median absolute deviation. MRI measures do not include arms and head. CPAI, Cambridge physical activity index; HC, hip circumference.

5.2.2.2 C-JAMP: Genetic effects on obesity

First, results from the genetic association analysis with log(SAT) and $\frac{\text{SAT}}{\text{TAT}}$ are reported. In total, 482,507 SNVs were tested for their association with either trait. Using a significance threshold of $\alpha = 1.0 \times 10^{-7}$ (i.e., Bonferroni-correction for multiple testing), both C-JAMP and univariate regression were not able to identify any SNV associated with either trait. Also, if the analysis was restricted to the 304,457 SNVs with MAF> 0.01, none of the SNVs passed the respective Bonferroni-adjusted level. Finally, if the genotype-obesity associations were evaluated on the gene-level (as in section 4.2) instead of on the SNV level using C-JAMP,

also no genes were identified (using the Bonferroni correction to adjust for testing multiple SNVs within the gene and for testing multiple genes).

For a comparison of C-JAMP with univariate regression, Tables 5.7-5.8 show the 10 SNVs with smallest p-values under the C-JAMP model together with information about their genomic position, and with p-values under C-JAMP and univariate regression - with respect to testing genetic associations with log(SAT) (Table 5.7) and $\frac{\text{SAT}}{\text{TAT}}$ (Table 5.8). As a first observation, the p-values under C-JAMP tests are consistently smaller compared to p-values under univariate models, for almost all SNVs in Tables 5.7 and 5.8, and also in general for most other small p-values (see Figure 5.3 for a direct comparison of all p-values). In addition, it can be noted that almost all of the top SNVs using C-JAMP were also among the top SNVs using linear regression.

rsID	CHROM	POS	Gene	MAF	C-JAMP	LM
rs2514681	8	100213331	SPAG1	0.11	9.7×10^{-7}	1.4×10^{-5}
rs116280906	22	36257940	APOL1	0.03	6.6×10^{-6}	1.7×10^{-5}
–	1	29193089	MECR	0.03	1.1×10^{-5}	1.3×10^{-5}
–	20	29480155	ABBA01031664.1	0.03	1.5×10^{-5}	7.3×10^{-5}
rs8111684	19	48444069	GRIN2D, GRWD1	0.07	1.7×10^{-5}	7.0×10^{-5}
rs1284539	1	146994776	NBPF12	0.18	1.8×10^{-5}	9.0×10^{-6}
rs150506419	9	32987844	APTX	0.02	1.9×10^{-5}	3.3×10^{-5}
rs75033136	14	91871327	FBLN5	0.14	1.9×10^{-5}	1.0×10^{-4}
rs75525243	19	19527409	NDUFA13, YJEFN3	0.21	2.0×10^{-5}	2.9×10^{-5}
rs72829161	5	148875842	ADRB2, SH3TC2	0.06	2.2×10^{-5}	2.1×10^{-5}

Table 5.7: Top SNVs associated with log(SAT) using C-JAMP in the obesity study. Shown are the top 10 SNVs with smallest p-values (for testing $H_0 : \alpha_{XY} = 0$) using C-JAMP, with information about their genomic position and MAF, and the unadjusted raw p-values from single-marker tests using C-JAMP and LM. rsID is the reference ID of the SNV.

Regarding a biological interpretation of the C-JAMP results, it was observed that the top SNVs for body fat mass are very different from the top SNVs for body fat distribution, and all top hits with respect to SAT don't show any indication (even p-values<0.05) for an association with $\frac{\text{SAT}}{\text{TAT}}$, and vice versa. Even though none of the SNVs passed a genome-wide significance threshold, some of the SNVs in Table 5.7 and 5.8 could be interesting targets for follow-up analyses. The most direct link might be drawn for rs116280906 in the APOL1 gene, which encodes the high density apolipoprotein L1 and is involved in cholesterol and lipoprotein metabolic pathways. rs116280906 was positively associated with SAT mass ($\widehat{\alpha}_{XY} = 1.48$, $\widehat{SE}_{\widehat{\alpha}_{XY}} = 0.33$, p-value=$6.6 \times 10^{-6}$) and further investigations indicated that the SAT gene expression of APOL1 was positively correlated with SAT mass (Pearson correlation $r = 0.22$, p-value=0.005) while the SNV showed only a very weak positive association with APOL1 gene expression ($\widehat{\beta} = 0.55$, $\widehat{SE}_{\widehat{\beta}} = 0.30$, p-value=0.06 from a linear regression of the gene expression conditional on the SNV and covariates). This provides interesting hints that rs116280906 and the APOL1 gene can be interesting candidates for follow-up studies. Other interesting markers could be rs6838636 in MAN2B2 which is involved in saccharide metabolic

pathways, and rs1048365 which is in vicinity of the VGF gene involved in insulin pathways and generation of precursor metabolites.

rsID	CHROM	POS	Gene	MAF	C-JAMP	LM
rs147163321	19	40944788	CYP2B7P	0.01	4.3×10^{-7}	7.6×10^{-6}
rs6838636	4	6619461	MAN2B2	0.01	1.5×10^{-6}	6.6×10^{-6}
rs1048365	7	101161149	AP1S1, VGF	0.11	2.2×10^{-6}	5.0×10^{-6}
rs2303710	5	6602555	NSUN2	0.21	2.4×10^{-6}	1.8×10^{-5}
rs3747213	22	44186256	PARVG	0.26	2.6×10^{-6}	4.7×10^{-5}
rs60320548	4	187436265	RP11-91J3.3	0.40	2.7×10^{-6}	2.1×10^{-5}
–	9	110172513	AKAP2	0.05	3.1×10^{-6}	4.5×10^{-5}
rs9912684	17	82915191	TBCD	0.08	3.4×10^{-6}	2.4×10^{-5}
rs2368558	14	103764181	PPP1R13B	0.07	3.5×10^{-6}	2.9×10^{-5}
rs7374776	3	196555215	WDR53	0.06	3.5×10^{-6}	4.1×10^{-5}

Table 5.8: Top SNVs associated with $\frac{\text{SAT}}{\text{TAT}}$ using C-JAMP in the obesity study. Shown are the top 10 SNVs with smallest p-values (for testing $H_0 : \beta_{XY} = 0$) using C-JAMP, with information about their genomic position and MAF, and the unadjusted raw p-values from single-marker tests using C-JAMP and LM. rsID is the reference ID of the SNV.

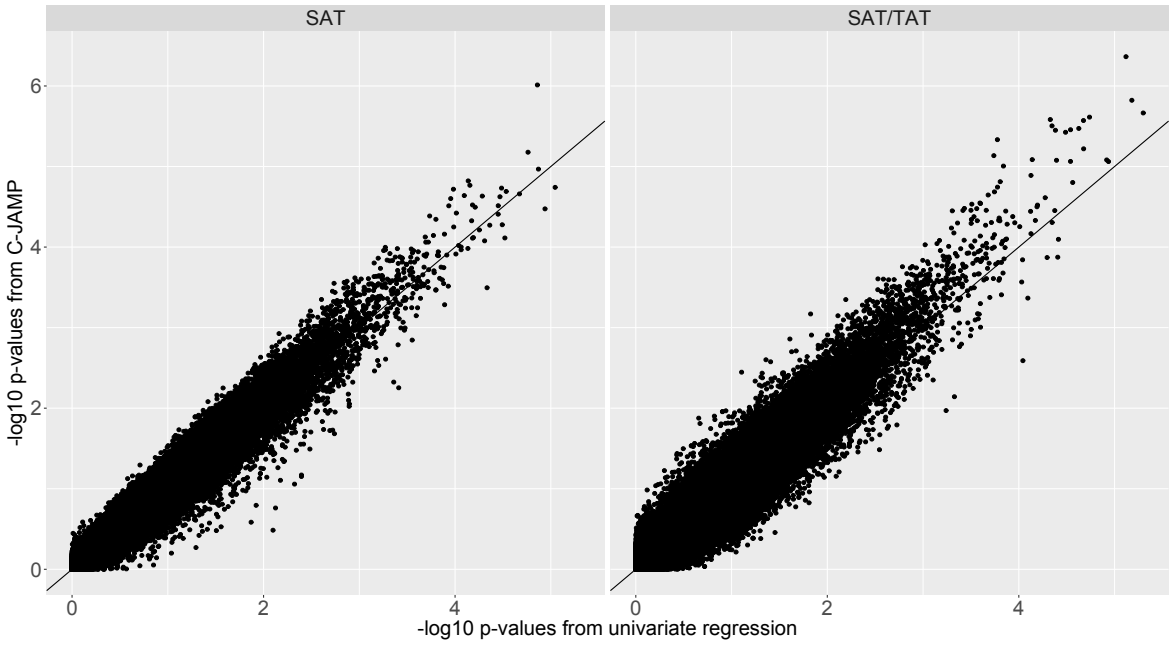


Figure 5.3: Scatterplots of p-values from C-JAMP versus linear regression from the genetic association analysis with obesity traits. P-values are on a $-\log_{10}$ scale from C-JAMP models of SAT and $\frac{\text{SAT}}{\text{TAT}}$ conditional on SNVs and covariates with grey dashed lines at the Bonferroni-adjusted significance threshold, for the 304,457 out of the 482,507 SNVs with MAF>0.01.

5.2.2.3 C-JAMP: Transcriptomic effects on obesity

Next, results from the transcriptomic association analysis with log(SAT) and $\frac{\text{SAT}}{\text{TAT}}$ are reported. In total, 30,917 genes were each tested for their association with either trait. In a first step, the results of analyses based on C-JAMP and univariate regression are compared. Using a transcriptome-wide significance threshold of $\alpha = 3.2 \times 10^{-5}$ (Bonferroni-adjusted $\alpha = 0.05$), C-JAMP identified more associated genes (441 with respect to SAT and 225 with respect to $\frac{\text{SAT}}{\text{TAT}}$) as compared to univariate regression (410 with respect to SAT and 121 genes with respect to $\frac{\text{SAT}}{\text{TAT}}$). Similarly to the genetic association analyses, C-JAMP and linear regression yield the same top hits in a very similar order, but C-JAMP provides consistently much smaller p-values. This can be seen in Tables 5.9-5.10 as well as in Figure 5.4, and confirms the previous results that C-JAMP yields higher power to identify molecular loci compared to univariate analyses.

CHROM	Gene	C-JAMP	LM	GO annotation
3	GLB1	4.7×10^{-16}	3.9×10^{-14}	Metabolic & catabolic processes, inflammation
X	IRAK1	1.4×10^{-15}	1.2×10^{-13}	Signaling pathways, transcription regulation
1	CDC20	2.9×10^{-15}	3.0×10^{-13}	Cellular processes
22	SLC7A4	1.1×10^{-14}	2.5×10^{-12}	Cell transport
4	PALLD	1.6×10^{-14}	1.8×10^{-12}	Cellular processes & transport
22	P2RX6	2.0×10^{-14}	9.5×10^{-13}	Signaling pathways, cellular processes
17	ABCC3	7.0×10^{-14}	4.9×10^{-13}	Cell transport
12	SPX	7.2×10^{-14}	1.7×10^{-12}	Regulation of cardiometabolic processes appetite, blood pressure
17	RASL10B	1.4×10^{-13}	8.6×10^{-13}	Signal transduction, regulation of cardiometabolic processes blood pressure, hormones
4	UCHL1	1.7×10^{-13}	6.5×10^{-12}	Cell transport, neurological processes

Table 5.9: Top genes whose gene expression is associated with log(SAT) using C-JAMP in the obesity study. Shown are the top 10 genes with smallest p-values (for testing $H_0 : \beta_{GY} = 0$) using C-JAMP, with information about their genomic position, and the p-values using C-JAMP and LM. 'GO Annotation' is a summary of the GO terms associated with the gene.

CHROM	Gene	C-JAMP	LM	GO annotation
3	ALDH1L1-AS2	4.8×10^{-13}	6.9×10^{-10}	–
1	CDKN2C	8.2×10^{-13}	1.4×10^{-10}	Cell transport
15	TNFAIP8L3	1.3×10^{-12}	1.8×10^{-10}	Signaling pathways, metabolic processes
3	ALDH1L1	2.8×10^{-12}	8.2×10^{-10}	Metabolic and catabolic processes
7	GJC3	2.4×10^{-11}	1.6×10^{-09}	Neurological & sensory processes
11	PHLDA2	2.4×10^{-11}	2.3×10^{-10}	Cell transport, metabolic processes
7	SERPINE1	5.2×10^{-11}	1.6×10^{-08}	Cell transport, inflammation
16	LDHD	6.7×10^{-11}	1.1×10^{-09}	Metabolic processes
8	TRIM55	8.0×10^{-11}	2.2×10^{-09}	Signal transduction
7	AZGP1	8.4×10^{-11}	5.5×10^{-09}	Transport of cells & molecules

Table 5.10: Top genes whose gene expression is associated with $\frac{\text{SAT}}{\text{TAT}}$ using C-JAMP in the obesity study. Shown are the top 10 genes with smallest p-values (for testing $H_0 : \beta_{GY} = 0$) using C-JAMP, with information about their genomic position, and the p-values using C-JAMP and LM. 'GO Annotation' is a summary of the GO terms associated with the gene.

For a biological interpretation of the results, Tables 5.9-5.10 also show a summary of the GO terms associated with each gene. While they provide a first idea about the role of each gene, some of the genes have not been studied extensively, so that these descriptions should be interpreted with caution. For a more robust picture, the results of C-JAMP were followed-up with a GO-term enrichment analysis for SAT and $\frac{\text{SAT}}{\text{TAT}}$, respectively, based on the 441 genes associated with SAT and 225 genes associated with $\frac{\text{SAT}}{\text{TAT}}$. The analysis was based on GO terms related to biological processes, and three different tests for enrichment (Fisher's exact test, classic and elim KS) were computed. As the main result of this analysis, the top ranked GO terms indicate that the genes associated with SAT mass are mostly related to processes and reactions of the immune system, inflammation, and cell transport. On the other hand, the genes associated with the relative amount of SAT (i.e., $\frac{\text{SAT}}{\text{TAT}}$) are mostly related to metabolic and catabolic processes, and immune system processes and inflammation as well. This matches the results for the top ranked genes but provides a more robust picture.

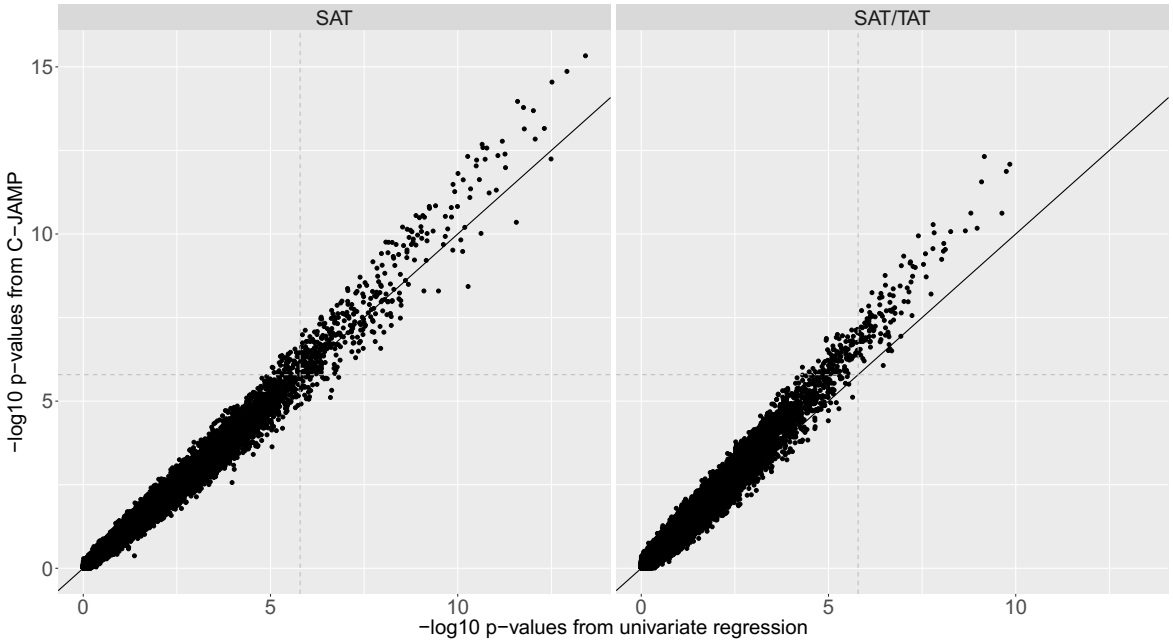


Figure 5.4: Scatterplots of p-values from C-JAMP versus linear regression from the transcriptomic association analysis with obesity traits. P-values are on a $-\log_{10}$ scale from C-JAMP models of SAT and $\frac{\text{SAT}}{\text{TAT}}$ conditional on gene expression and covariates with grey dashed lines at the Bonferroni-adjusted significance threshold $\alpha = 3.2 \times 10^{-5}$.

In a more detailed discussion of the results, it can be seen that while some of these biological processes are identified by all three tests, there are marked differences between the top-ranked genes of each test. For an illustration, Table 5.11 shows an extract of the results in form of the top 20 GO terms that are enriched in the 441 SAT-genes and 225 $\frac{\text{SAT}}{\text{TAT}}$-genes, respectively, based on the classic KS-test. For the full results, see Supplementary Tables A.28-A.33.

SAT			$\frac{\text{SAT}}{\text{TAT}}$		
#	GO ID	GO term name	#	GO ID	GO term name
1	0002376	immune system process	1	0044710	single-organism metabolic process
2	0006955	immune response	2	0044281	small molecule metabolic process
3	0002682	regulation of immune system process	3	0044699	single-organism process
4	0001775	cell activation	4	0044763	single-organism cellular process
5	0045321	leukocyte activation	5	0002376	immune system process
6	0002252	immune effector process	6	0001775	cell activation
7	0002684	positive regulation of immune system process	7	0006955	immune response
8	0016192	vesicle-mediated transport	8	0045321	leukocyte activation
9	0050776	regulation of immune response	9	0055114	oxidation-reduction process
10	0051234	establishment of localization	10	0006082	organic acid metabolic process
11	0046649	lymphocyte activation	11	0002684	positive regulation of immune system process
12	0051179	localization	12	0043436	oxoacid metabolic process
13	0050778	positive regulation of immune response	13	0002682	regulation of immune system process
14	0006810	transport	14	0032787	monocarboxylic acid metabolic process
15	0002263	cell activation involved in immune response	15	1902578	single-organism localization
16	0042110	T cell activation	16	0006954	inflammatory response
17	0002366	leukocyte activation involved in immune response	17	0019752	carboxylic acid metabolic process
18	0019882	antigen processing and presentation	18	0044765	single-organism transport
19	0048002	antigen processing & presentation of peptide antigen	19	0006091	generation of precursor metabolites and energy
20	0050896	response to stimulus	20	0006629	lipid metabolic process

Table 5.11: Results of the GO term enrichment analysis in the obesity study, of the 441 genes associated with SAT (left panel) and 225 genes associated with $\frac{\text{SAT}}{\text{TAT}}$ (right panel). Shown are the top 20 GO terms with smallest p-values based on the classic KS test.

5.2.2.4 CIEE: Direct genetic effects on obesity

In a follow-up to the C-JAMP analyses in the previous sections, CIEE was used to test direct genetic effects of the 15,895 SNVs in the identified 607 genes on both obesity traits, while removing the indirect effect through gene expression. First checks showed that CIEE did not provide valid inference for singletons and doubletons, and hence the analysis was restricted to the 10,238 SNVs with at least 4 observed minor alleles (i.e., MAF$\geq$ 0.01). Using the genome-wide threshold of $\alpha = 1.0 \times 10^{-7}$, CIEE identified 17 SNVs associated with SAT and 52 SNVs associated with $\frac{\text{SAT}}{\text{TAT}}$. Of these SNVs, 7 SNVs (1 for SAT and 6 for $\frac{\text{SAT}}{\text{TAT}}$) had MAF$\geq$ 0.025 (i.e., at least 10 observed minor alleles). An overview about these 7 SNVs with genomic information and results from CIEE as well as the MR in equation 5.2.7 is shown in Table 5.12. Regarding the traditional regression approaches MR and RR, both were not able to identify any associated SNVs using the genome-wide threshold of $\alpha = 1.0 \times 10^{-7}$ or the more liberal $\alpha = 3.1 \times 10^{-6}$ (corrected for testing 15,895 SNVs).

The results in Table 5.12 show that while CIEE and MR provide similar standard error estimates, the point estimates $\widehat{\alpha}_{XY}$ of CIEE are much larger. Based on the results of the simulation study in section 4.3 and assuming the correctness of the assumed DAG, it is suggested that there is indeed either an indirect genetic effect in opposing direction of the direct effect through gene expression, or that there is unmeasured confounding of the indirect effect, which leads to biased estimates of the direct effect with MR. Regarding a biological interpretation of the results, of particular interest might be the identified SNVs in GPT2 and ERAP1. GPT2 encodes the enzyme alanine aminotransferase 2 and is involved in the amino acid metabolism. In our study, the SNV in GPT2 showed a negative association with

SAT mass. While an unbiased estimation of the indirecte effect of the SNV on SAT mass through gene expression would require more elaborate methods, the following results can give some indication on the underlying biology and effects: The SAT gene expression of GPT2 was negatively correlated with SAT mass (Pearson correlation $r = -0.22$, p-value=0.001) and while the SNP was not associated with GPT2 gene expression ($\widehat{\beta} = -0.44$, $\widehat{SE}_{\widehat{\beta}} = 0.35$, p-value=0.21), the overall indirect effect might be positive and hence in opposite direction as the direct effect as estimated by CIEE ($\widehat{\alpha}_{XY} = -0.96$) - which caused MR to provide biased estimates. Another candidate of interest for follow-up investigations could be ERAP1, which encodes the endoplasmic reticulum aminopeptidase 1 and is involved in fat cell differentiation, peptide catabolic processes and blood pressure regulation.

Trait	SNV					$\widehat{\alpha}_{XY}$		$\widehat{SE}(\widehat{\alpha}_{XY})$		P-value	
	rsID	CHROM	POS	MAF	Gene	CIEE	MR	CIEE	MR	CIEE	MR
SAT	–	16	46924178	0.03	GPT2	-0.96	-0.63	0.18	0.29	8.1×10^{-8}	0.03
$\frac{\text{SAT}}{\text{TAT}}$	rs12646586	4	15251025	0.04	RP11-665G4.1	-0.98	-0.24	0.18	0.13	7.3×10^{-8}	0.07
	rs1460072	4	165873677	0.05	TLL1	0.88	0.07	0.15	0.15	1.4×10^{-9}	0.65
	rs72753967	5	41150324	0.04	C6	-0.93	-0.17	0.15	0.15	2.3×10^{-10}	0.24
	rs295670	5	58851926	0.04	RAB3C	0.85	0.28	0.13	0.14	2.8×10^{-11}	0.04
	rs249957	5	96785338	0.03	ERAP1	0.99	0.21	0.15	0.20	1.8×10^{-11}	0.29
	rs488492	11	79214799	0.11	TENM4	0.68	0.08	0.13	0.09	5.7×10^{-8}	0.37

Table 5.12: Results of the SNVs associated with SAT and $\frac{\text{SAT}}{\text{TAT}}$ using CIEE. Shown is information about their genomic position, coefficient and standard error estimates of the direct genetic effect and according p-values, from CIEE and from MR. rsID is the reference ID of the SNV.

Chapter 6

Discussion

6.1 Summary and data analysis recommendations

In this thesis, the two new statistical approaches C-JAMP and CIEE have been introduced, described regarding their statistical details and implementation, and evaluated for genetic association analyses incorporating multiple traits, hence meeting the goals set out in Aim 1. While C-JAMP allows investigating multiple traits in a joint model conditional on genetic variants and covariates which can improve the statistical power to identify associated genetic variants, CIEE allows removing indirect genetic effects through intermediate phenotypes in order to estimate and test direct genetic effects on a primary phenotype in a valid and robust manner. Both approaches provide an unbiased and efficient estimation of parameters and can be applied to different outcome variables analyzing multiple phenotypes and adjusting for multiple covariates, and have a clear interpretation of the estimated parameters. Aim 2 - to identify novel genetic and transcriptomic effects on cardiometabolic traits - has been met through the two empirical applications of C-JAMP and CIEE.

In extensive simulation studies as well as in empirical applications, C-JAMP and CIEE have been applied and evaluated for the analysis of common and rare genetic variants with quantitative phenotypes as well as time-to-event phenotypes. First, single-marker and multi-marker tests were compared for association analyses of rare variants, then C-JAMP was compared with the univariate SMT and MMTs as well as multivariate MMTs, and finally CIEE was compared to traditional and newly proposed approaches for the analysis of direct effects of common genetic variants. The results indicate that C-JAMP and CIEE (1) yield valid and efficient genetic effect estimates and can increase the power of hypothesis tests compared to separate univariate models, and (2) provide an improvement over existing methods for association analyses of rare genetic variants and for the identification of direct effects of common genetic variants, hence supporting the proposed thesis in section 1.1. The results can be summarized and combined with established results from the literature as shown in boxes 6.1-6.2 in order to form guidelines for practical analyses. A more detailed discussion of the two proposed methods, of the results from the simulation studies comparing the different methods, and of the empirical findings from the application studies are given in the following.

Association studies with rare genetic variants

1. **Association analyses of quantitative traits generally yield higher power compared to association analyses of dichotomized traits (section A.4.1.3).**
2. **Association analyses based on joint models of multiple phenotypes can improve the power compared to separate analyses (sections 4.2, A.4.2).**
3. **For the analysis of dichotomous phenotypes, many univariate SMTs have inflated type I error and univariate MMTs have higher power compared to univariate SMTs (see references in section 2.2.1 and section A.4.1.3).**
4. **For the analysis of quantitative phenotypes:**
   (a) **Univariate SMTs lead to more powerful association tests for identifying causal genes than univariate MMTs when the effect sizes of causal variants in the gene are large, and less powerful tests when causal variants have small effect sizes. For moderate effect sizes, if at most 5% or 10% rare SNVs in a gene are causal, then the univariate SMT has higher power when the sample size is larger, and slightly less power for a smaller sample size (sections 4.1, A.4.1).**
   (b) **Many of the proposed multivariate MMTs investigated here have invalid type I errors and should not be used for rare variant analyses (section 4.2).**
   (c) **C-JAMP provides valid estimation and inference in the analysis of rare variants - given that the dependence between traits is correctly modeled and an appropriate test statistic adjustment is made for the analysis of very rare variants with traits of high dependence (section 4.2). C-JAMP has the highest power of all investigated approaches if less than 50% of SNVs are causal and the traits have high dependence.**

Box 6.1: Summary for association studies with rare genetic variants.

6.2 Single-marker tests versus multi-marker tests

In the comparison of SMTs and MMTs in this thesis, the focus was on the association analysis of rare variants with quantitative traits. First, it was shown that standard restricted maximum likelihood estimation in a single-marker approach provides unbiased estimates and single-marker Wald-type t-tests allow valid statistical inference even for very rare SNVs. Next, the power of the SMT and MMTs was compared for identifying a gene that includes causal

rare variants, and it was shown that SMTs or MMTs can be more powerful depending on the scenario. More specifically, SMTs lead to more powerful association tests for identifying causal genes than MMTs when the effect sizes of causal variants in the gene are large, and less powerful tests when causal variants have small effect sizes. For moderate effect sizes, whether SMTs or MMTs have higher power depends on the sample size and percentage of causal SNVs. When at most 5-10% SNVs in a gene are causal, which might be the most realistic scenario, the SMT has higher power when the sample size is larger, and slightly less power for a smaller sample size. These results seem to hold consistently for different gene sizes, different nominal α levels, different proportions of SNVs with positive/negative effect on the trait, and different corrections for multiple testing of SNVs in a gene (Bonferroni and BH correction) with SMTs. In the empirical data analysis of the GAW19 data in section A.4.1.4, the SMT provided smaller p-values compared to the MMTs in the testing of previously identified blood pressure genes.

Estimation and testing of direct genetic effects

1. **All approaches rely on the assumption that the assumed directed acyclic graph model is correct. Furthermore (see sections 4.3, A.4.3),**
2. **CIEE provides valid estimation of the direct genetic effect and its standard error, as well as valid testing of the direct genetic effect in all considered scenarios, also when there is unmeasured confounding of the indirect genetic effect through observed factors - for both the analysis of quantitative and time-to-event primary traits.**
3. **Its direct genetic effect estimates coincide with the estimates from the G-estimation method for quantitative primary traits (Vansteelandt et al., 2009), which also provides valid testing but does not provide standard error estimates of the effect.**
4. **The G-estimation method for time-to-event primary traits (Lipman et al., 2011) provides biased effect estimates and inflated type I errors in hypothesis testing.**
5. **Traditional regression and structural equation modeling methods provide valid parameter estimates and hypothesis tests of direct genetic effects, if there is no unmeasured confounding of the indirect effect.**
6. **The power of all approaches in hypothesis tests of the direct genetic effect is similar in all scenarios where they provide valid testing.**

Box 6.2: Summary for the estimation and testing of direct genetic effects.

For any gene-level investigation, a very important question is which SNVs should be grouped together in a test. Including more true causal SNVs by extending the region under consideration or by analyzing both common and rare SNVs can increase the power. At the

same time, including more non-causal SNVs adds noise and decreases the power. Accordingly, in the simulation study, the power of all approaches increased when the analysis was restricted to rare SNVs, since only rare SNVs were chosen as causal SNVs in the data generation. The same was observed in the real data analysis - but only consistently for the SMT - where the reason might have been as well that there were not any common causal SNVs. The data application also showed that while the SMT, SKAT, and SKAT-O are quite robust against which sets of SNVs are investigated, the burden test is highly sensitive and can lead to vastly different results.

It should be noted that in the simulation study, the data was generated from the same model as in Lee et al. (2012) where the effect size of a SNV depends on the MAF of SNVs through a specified function (see Table 4.1). Since the weighted kernel functions in SKAT and SKAT-O are dependent on the MAF of SNVs as well, this gives SKAT and SKAT-O an advantage and allowed evaluating and comparing the SMT in situations where the MMTs have high power. Therefore, in real applications, the power of SMTs might compare even more favorably to these MMTs.

The proposition to use SMTs for the analysis of rare variants is based on the fact that maximum likelihood estimation returns valid estimates under the linear regression model, even for the analysis of singletons and doubletons. Importantly, the strategy to include singletons and doubletons in the analysis can substantially improve the power of SMTs. Of course, this does not alleviate the need to check for genotyping errors and other biases, which are especially relevant to singletons and doubletons. Also, while all investigated approaches rely on normality assumptions, the Wald-type t-test (SMT) was much more vulnerable to a distribution misspecification than the MMTs. When the residuals in the genetic association test are not normally distributed for one or multiple SNVs (or when effects follow another genetic model), adapting the test statistic is warranted.

Further simulation studies (data not shown) confirmed that for binary outcomes, standard statistical tests such as Fisher's exact test, logistic regression likelihood ratio tests, Wald tests, and Score tests, should not be directly used for testing singletons and doubletons since they do not allow valid inference. Therefore, for each test, all assumptions including the (asymptotic) distribution assumptions and empirical type I errors have to be assessed before the analysis, and significant test results have to be cautiously investigated, especially if they are based on singletons or doubletons.

Secondary results of the simulation study provide suggestions regarding the study design of rare variant association studies. It was observed that the power to identify causal low-frequency SNVs is almost always much smaller compared to the power to identify gene regions encompassing causal rare SNVs (cf. Ladouceur et al., 2012), except when very few SNVs in a gene are causal, have higher MAF (i.e., MAF between 0.005 and 0.03) and higher effect sizes. This should be kept in mind but not be considered as a power comparison between MMTs and SMTs. Hence, the power of rare variant tests depends highly on which hypothesis is tested, the number of causal SNVs, their effect size, and whether the ultimate goal is to identify a causal gene or a causal SNV. The comparison between SNV-level and gene-level association tests might also depend on the ratio of the number of genes to the number of SNVs, which could be further investigated in future studies. Regarding MMTs, the results

come with a loss in information because the exact position of causal variants is not revealed, and to my knowledge, the currently available methodology does not allow to obtain power estimates for identifying a causal SNV. If the focus of the analysis is on identifying causal rare SNVs instead of genes, SMTs can be used in a genome-wide scan for variants, or alternatively SMTs/MMTs can be used in the first stage to identify target genes and followed up by SMTs of variants within these genes. Comparisons between such 1-stage and 2-stage approaches were not the aim of this study, and could be an interesting avenue for future research by incorporating correct adjustments for multiple testing, the 2-stage testing and biases such as the winner's curse (Sun et al., 2011). For choosing the most powerful statistical test for specific situations in genetic association studies of rare variants, the locus of interest (gene or variant) should be specified first, followed then by a comparison of different statistical tests.

These results add more details and shed a light on the comparison of single-marker and multi-marker tests for quantitative traits, and provide clear suggestions under which assumptions MMTs should be used, and when a simple SMT is favorable. In contrast to popular practice in epidemiological and medical studies, the additional results also highlight that quantitative traits should be analyzed whenever possible instead of binary traits to maximize the power of association tests.

6.3 C-JAMP

In the simulation studies and empirical applications, C-JAMP provided highest power compared to all other - multivariate and univariate - approaches in all scenarios with two exceptions. As first exception, SKAT-O had slightly higher power compared to C-JAMP when 50% of all rare SNVs were causal and they had moderate or small effect sizes all in the same direction. Second, the power of C-JAMP was similar to or slightly lower compared to the aSPUset-Score test, when the dependence between traits was $\tau = 0.2$, there were 20% causal SNVs in a gene and effect sizes were high or there were 50% causal SNVs in a gene. Interestingly, most investigated alternative approaches based on multivariate models yielded inflated type I errors. Only two tests, aSPUset and aSPUset-Score had acceptable type I errors (even though aSPUset-Score also showed a slight inflation), but had much smaller power compared to C-JAMP in almost all considered scenarios. For a comparison of C-JAMP with standard multivariate regression approaches, it can be noted that the aSPUset-Score test includes the GEE-Score test (which is equivalent to the MANOVA), so that these approaches have been implicitely incorporated in the comparison. Earlier work (Konigorski et al., 2014) also showed that C-JAMP provided a much better model fit compared to a bivariate linear model.

A downside of aSPUset and aSPUset-Score, and of MMTs in general, is that they rely on a multitude of assumptions (e.g., many SNVs in a gene are causal) and parameters (e.g., the choice of power sets), and even an integration of many different subtests does not yield tests with "optimal" power, as shown by the results of the simulation studies. On the other hand, SMTs rely on much fewer assumptions. In addition, the aSPUset and aSPUset-Score tests are computationally very intensive and not suitable for genome-wide analyses, where evidence in form of p-values smaller than 10^{-5} of 10^{-8} is needed. That is, since permutation or simulation approaches are employed to derive p-values of aSPUset and aSPUset-Score

empirically, 10^k permutations are needed to be able to obtain p-values that are potentially as small as $\frac{1}{10^k+1}$.

C-JAMP seems to be well-suited for many applications since it can fit different dependence structures between two (or more) traits, does not lead to a power decrease when SNVs are only associated with one trait - in contrary to omnibus multi-degree-of-freedom tests - and is computationally fast enough to be employed on a genome-wide scale. Furthermore, C-JAMP provides flexibility to use different genetic models for the effect of each SNV and to use any marginal distribution of a given trait conditional on the SNV and other factors. As further practical arguments for integrating multiple traits in the analysis, in many situations, multiple traits of a phenotype with moderate or strong dependence (for example, Kendall's τ between 0.5 and 0.8) can easily be obtained, since they are often measured concurrently. For example, systolic and diastolic blood pressure or different measures of body fat mass and body fat distribution are assessed in most investigations into blood pressure or obesity, and they often have high dependence. Consequently, the application studies in chapter 5 illustrated that C-JAMP identified more candidate SNVs and genes with potential function in blood pressure and obesity genetics.

Regarding the application of C-JAMP to genome-wide analyses, it is fast enough to analyze millions of SNVs within days by using a computing cluster, but is still much slower compared to standard regression approaches, especially if large sample sizes and a large number of predictors are analyzed. The computational expensive part of C-JAMP consists of the likelihood maximization (and looping over different starting values in case of non-convergence) which is currently done using the BFGS algorithm in the `optimx` R package. This algorithm was chosen since it provided the most robust optimization compared to all other methods implements in the `optimx` package (and therefore available in `optim` and `nlm`) in addition to speed. The focus was primarily set on robustness correctness of parameter estimates, convergence, fulfilment of the KKT conditions, estimation of the Hessian matrix for the parameter estimates in order to obtain standard error estimates. Further problems can arise in the optimization with BFGS when the different predictors have different scaling, which was hence built into the `cjamp()` R function. Future work can investigate approaches to decrease the running time. One avenue can be to investigate other optimization algorithms that are available in R and also more efficient implementations of the optimization algorithms (Nash, 2014). Alternatively, restricting the parameter space could help, even if preliminary results from an ongoing investigations indicate a loss of robustness. Finally, the computationally expensive part could be accelerated using graphics processing units (GPUs), for example using CUDA libraries (for examples, see Buckner et al., 2010; Suchard et al., 2010).

As second caveat for the practical application of C-JAMP, an appropriate copula function has to be chosen according to the dependence structure between the traits. The 2-parameter copula function used in this study covers a wide range of possible dependencies suitable for most applications, and can be complemented with the choice of a Gaussian copula. Third, in the analysis of very rare SNVs with highly dependent traits, an adjustment of the Wald test statistic was necessary to correct for inflated/deflated test statistics. While the approach is generalizable and can be used in practical investigations as well, future work can focus on the derivation of the theoretical distribution of the test statistic or also focus on an

unbiased estimation of parameters and standard errors on the boundary of the parameter space. Alternatively, SNVs with very low frequency can be deleted before the analysis to avoid any problems, but this might lead to a small power decrease in gene-based tests, as described in section A.4.1.1. Fourth, in the situation when there is a strong dependence between the traits, but the SNV of interest is associated with only one of the traits, then there is an indication for inflated type I errors for testing the association with the non-associated trait. Even though this scenario might be unlikely in practice when two traits with common biological underpinnings are investigated, comparing the results of C-JAMP with univariate models can serve as a check. In future work, the implementation of C-JAMP can also incorporate more than 2 phenotypes to allow building more complex models. It could be expected that this allows to further increase the power, but at the same time a saturation effect of the power increase caused by the increasing number of estimated marginal parameters could come into play.

6.4 CIEE

In CIEE, I propose to use the estimating equation method to estimate direct genetic effects on a primary phenotype, adjusting for indirect effects through intermediate phenotypes that can also be influenced by measured or unmeasured confounding factors. Multiple influencing factors and multiple intermediate phenotypes can be included in the model. CIEE yields consistent estimates for the direct effect and provides accurate standard error estimates, even when there is unmeasured confounding of the indirect effect through measured factors. For the analysis of time-to-event data subject to censoring, CIEE includes a new approach for the removal of the intermediate effect. Due to the use of the established theory of estimating equations, the approach can be extended to different error distributions. At the same time, the use of robust sandwich standard error estimates also provides valid inference if the error distribution is misspecified, as shown in the simulation study. Of note, when analyzing quantitative traits, CIEE yields estimates equivalent to the LS estimates under the corresponding models, which do not rely on any distribution assumption. Therefore, the resulting direct effect estimate is accurate even if the distribution assumption is not satisfied.

Applying CIEE to genetic association studies can both identify genetic variants that would be missed by traditional analyses, and can prevent false positive results - depending on whether the indirect genetic effect with unmeasured confounders is in the same or opposite direction of the direct effect. This was illustrated in the real-data applications in chapter 5, where novel candidate markers were identified for systolic blood pressure, body fat mass, and body fat distribution which were missed by traditional regression approaches. The results of the simulation study also provided a detailed analysis when other methods provide valid estimation and testing, and when they should not be used. Standard multiple regression approaches were valid in all scenarios as long as there was no unmeasured confounding of the indirect genetic effect. For example, they also provided valid inference when there was measured confounding of the indirect genetic effect - which is in contrast to some claims in the literature (Goetgeluk et al., 2008). The genetic effect estimates obtained from SEM were also affected by unmeasured confounding of the indirect genetic effect, which exemplifies that

SEM is highly dependent on the correctness of the assumed paths and edges and may lead to biased estimates otherwise. Finally, the sequential G-estimation method (Vansteelandt et al., 2009) provides equally valid testing compared to CIEE for the analysis of quantitative traits, but the G-estimation approach proposed by Lipman et al. (2011) for the analysis of primary time-to-event phenotype is not able to remove the effect of intermediate phenotypes leading to biased direct effect estimates and invalid testing. In addition, the sequential G-estimation methods do not provide a standard error estimate of the estimated direct effect.

For an application of CIEE and any other model to the analysis of DAG models, it should be noted that despite the robustness properties of CIEE, there are still some assumptions that are required for valid testing and estimation. One assumption is that there is no unmeasured confounding of the direct genetic effect (i.e., factors both affecting the genetic marker and primary phenotype). For genetic association studies, this assumption seems plausible and if any such factors (e.g., population stratification) were present, they could be controlled for in an initial step or by considering them as covariates. Furthermore, an a priori choice of relevant intermediate variables and influencing factors (i.e., distinction between K, L) is important. Next, while CIEE and the G-estimation methods are robust against unmeasured confounding of the indirect effect through measured factors, they lead to biased point estimates and inflated type I errors similar to traditional approaches if there is direct unmeasured confounding of the indirect effect (e.g., if U affects K directly and not only through L) so that the DAG is misspecified. Finally, additional results in the application of CIEE in section 5.2 showed that the large-sample Wald-type tests can lead to inflated type I errors when very rare SNVs are analyzed. This could be further investigated in future work. Also, future extensions of the implementation can incorporate multiple intermediate phenotypes and multiple influencing factors.

I believe that the application of CIEE to association studies in genetic epidemiology and other biomedical fields can provide new insights about direct effects of variables. In addition, future extensions of CIEE including multiple primary phenotypes in the analysis can provide further possibilities to build more complex and realistic models.

6.5 Empirical findings from cardiometabolic association studies

In the real-data applications, all analyses confirmed that C-JAMP and CIEE can identify candidate markers that would not be identified by traditional approaches. In addition, the identified candidate markers can provide novel insights and help to generate new hypotheses, in particular about markers and genes involved in obesity biology.

In the analyses of blood pressure in the GAW19 data, while none of the identified SNVs reached genome-wide significance (i.e., had a p-value smaller than 10^{-8}), both C-JAMP and CIEE yielded some SNVs with suggestive significance that had an adjusted p-valued smaller than 0.05 when using a Bonferroni-correction for multiple testing. C-JAMP identified 5 SNVs in or upstream of the CEACAM5 gene, and CIEE hinted the potential role of a new genetic locus upstream of the IL27RA gene, which is involved in anti-inflammatory processes and immune response (Hunter & Kastelein, 2012). These SNVs would have been missed if a

traditional regression analysis was performed.

In the second empirical study of obesity traits, a preliminary analysis investigated to what extent plasma adipokine concentrations can be predicted based on MRI-derived measures of AT mass of different body compartments, including SAT and VAT, and by SAT gene expression (section A.4.5). Relevant for the main investigation, the results showed that absolute body fat mass measures as well as relative body fat measures (which assess body fat distribution) showed similar correlations with plasma levels, so that SAT and $\frac{\text{SAT}}{\text{TAT}}$ were used in the main analyses described in section 5.2. In the genetic association analyses of total genetic effects on obesity with C-JAMP, none of the tested SNVs reached genome-wide significance for either trait, but there were several SNVs with p-values smaller than 10^{-5}. Most interesting for follow-up studies and replication analyses with a larger sample size might be rs116280906 in the APOL1 gene, which encodes the high density apolipoprotein L1 bound to HDL and is involved in cholesterol and lipoprotein metabolic pathways. To date, the APOL1 gene has mostly been investigated for its association with kidney disease (Genovese et al., 2010; Dummer et al., 2015) but has also been linked to obesity and other chronic diseases (Hu et al., 2012; Madhavan & O'Toole, 2014). Since in our study, an association between APOL1 gene expression and SAT mass was not observed, the marker rs116280906, the APOL1 gene and the respective apolipoprotein L1 seem interesting candidates for further investigations regarding their role in obesity. Other interesting markers could be rs6838636 in MAN2B2 which is involved in saccharide metabolic pathways, and rs1048365 which is in vicinity of the VGF gene involved in insulin pathways and the generation of precursor metabolites.

In the genetic association analyses of direct genetic effects on obesity controlling for indirect effects through gene expression with CIEE, 1 new locus was identified for SAT and 6 new loci were identified for $\frac{\text{SAT}}{\text{TAT}}$ (when restricting the analysis to SNVs with MAF ≥ 0.025; and 17, 52 SNVs were identified when restricting to SNVs with MAF ≥ 0.01, respectively), all with p-values smaller than 10^{-7} and 4 of them with p-values smaller than 10^{-8}. Of particular interest might be the identified SNVs in GPT2 and ERAP1. GPT2 encodes the enzyme alanine aminotransferase 2 and is involved in the amino acid metabolism (synthesis of pyruvate from L-alanine) and recent animal studies suggest an association GPT2 with obesity (McCurdy et al., 2016). While hypotheses about its role in obesity are difficult to make without a proper estimation of the indirect effect, the results described in section 5.2 suggest that the SNV in GPT2 has an overall negative (protective) effect on SAT mass, with a positive (detrimental) indirect effect through GPT2 gene expression and a larger "direct" protective effect (through any other biological processes which are here summarized in the "direct" effect). For biological hypotheses, the "indirect" genetic effect through gene expression or any other specific intermediate phenotypes can often constitute the more interesting pathways explaining the effect of genetic variants. Therefore, the development of new statistical methods complementing CIEE for an unbiased estimation of the indirect effect can help in the interpretation in future studies. A second candidate marker of interest for follow-up investigations identified in the CIEE analysis could be rs249957 in the ERAP1 gene, which encodes the endoplasmic reticulum aminopeptidase 1, is involved in fat cell differentiation, peptide catabolic processes and blood pressure regulation.

In the transcriptome-wide association analysis of SAT gene expression with SAT mass and $\frac{\text{SAT}}{\text{TAT}}$, the results of the GO-term enrichment analysis based on the 607 identified genes whose expression levels were associated with obesity suggest that different pathways are associated with fat mass and fat distribution. While SAT mass was mostly related to processes and reactions of the immune system, inflammation and cell transport, $\frac{\text{SAT}}{\text{TAT}}$ was mostly related to metabolic and catabolic processes, and immune system processes and inflammation as well. For an interpretation of the results, the interplay between the different cells present in the adipose tissue (adipocytes, innate cells such as macrophages, and adaptive immune cells) can be considered (Huh et al., 2014). When searching for genes associated with fat mass, it can be expected that many immunological processes and pathways are identified. An investigation of body fat distribution in the form of the ratio $\frac{\text{SAT}}{\text{TAT}}$, however, might adjust for these immunological cells and point to pathways specific for the SAT. This might have been why metabolic and catabolic pathways were more present with a higher frequency in genes associated with $\frac{\text{SAT}}{\text{TAT}}$ - and also why fewer genes are associated with $\frac{\text{SAT}}{\text{TAT}}$ since parts of the variance are removed.

Integrating the results of the genetic and transcriptomic association analyses, all results indicate the role of pathways including metabolic and catabolic processes in obesity and suggest specific candidates for relevant molecular processes. The fact that none of the SNVs with a direct genetic effect on SAT and $\frac{\text{SAT}}{\text{TAT}}$ identified by CIEE showed a statistically significant total genetic effect with obesity suggests that they act as eQTLs on the gene expression levels which in turn show an association in the opposite direction with obesity. For a more complete picture, a transcriptome-wide search for eQTL in cis and trans could provide helpful information.

These empirical results strongly support that including multiple molecular levels - here genetic and transcriptomic (and plasma concentrations in the preliminary study) - in the analysis together with multiple phenotypes can lead to the identification of candidate markers and hypotheses about interesting pathways, even if only rather small sample sizes are available for the analysis, if appropriate and powerful statistical methods are used. Including different traits (such as fat mass and fat distribution) can also conceptionally be thought to represent the different facets and possible pathways of the obesity phenotype. As a possible extension of the analyses, it would be interesting to include molecular measures on additional levels (e.g., epigenetic factors), or incorporate single-cell sequencing analyses (Vitak et al. 2017) of the different cells present in the adipose tissue.

As limitations, the small sample size of the analyses likely prohibited to identify further candidate variants. Also, replication analyses of all identified markers have not yet been carried out. Even though extensive quality control checks have been performed on all levels of the data generation, and all experiments were tested and performed according to protocol, different biases can have influenced the results. First of all, in the SAT biopsies, the amount of aspired tissue as well as the according RNA quality might depend on the obesity of the proband. In the RNA extraction, only total RNA was interrogated so that shorter RNA fragments and other regulatory elements were not assessed. This can be mostly seen as a limitation since their association with obesity traits could not be investigated, but it might also introduce some bias in the gene expression quantification if these short RNA molecules

are predominantly transcribed in some regions. On the other hand, the multiple quality control checks on all stages of the experiments and multiple normalization steps should limit these effects and the high association between gene expression levels based on RNA-Seq and PCR confirms the good validity of the RNA-Seq gene expression quantification. In the RNA sequencing, a comparison of the gene expression quantification based on a much higher sequencing depth indicated that the chosen depth (6 samples per lane) might be sufficient for the study purposes here. A comparison of the gene expression for 6 re-sequenced probes also indicated that the results are reliable. Furthermore, since the GO annotation of many different genes was integrated and all confirmed the same overall picture, this allows a higher certainty of the robustness of the results, and in the benefit of following-up the results in further analyses.

References

1. Ackermann M, Strimmer K (2009). A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, **10**: 47.
2. Aguiree F, Brown A, Cho NH, et al. (2013). *IDF diabetes atlas.* `www.diabetesatlas.org/key-messages.html`.
3. Aken BL, Ayling S, Barrell D, et al. (2016). The Ensembl gene annotation system. *Database: The Journal of Biological Databases and Curation*, **2016**: baw093.
4. Akram Z, Rahim ZH, Taiyeb-Ali TB, et al. (2017). Resistin as potential biomarker for chronic periodontitis: A systematic review and meta-analysis. *Arch Oral Biol*, **73**: 311-20.
5. Aleksandrova K, Boeing H, Jenab M, et al. (2012a). Leptin and soluble leptin receptor in risk of colorectal cancer in the European Prospective Investigation into Cancer and Nutrition cohort. *Cancer Res*, **72**: 5328-37.
6. Aleksandrova K, Boeing H, Jenab M, et al. (2012b). Total and high-molecular weight adiponectin and risk of colorectal cancer: the European Prospective Investigation into Cancer and Nutrition Study. *Carcinogenesis*, **33**: 1211-8.
7. Alexa A, Rahnenführer J, Lengauer T (2006). Improved scoring of functional groups from gene expression data by decorrelating go graph structure. *Bioinformatics*, **22**: 1600-7.
8. Alexa A, Rahnenführer J (2016). *topGO: enrichment analysis for Gene Ontology.* R package version 2.28.0.
9. Allaire JJ, Cheng J, Xie Y, et al. (2017). *rmarkdown: dynamic documents for R.* R package version 1.6. `https://CRAN.R-project.org/package=rmarkdown`.
10. Anders S, Pyl PT, Huber W (2015). HTSeq - a Python framework to work with high-throughput sequencing data. *Bioinformatics*, **31(2)**: 166-9.
11. Andrews S (2010). *FastQC: a quality control tool for high throughput sequence data.* Available at: `http://www.bioinformatics.babraham.ac.uk/projects/fastqc`.
12. Arnardottir ES, Maislin G, Jackson N, et al. (2013). The role of obesity, different fat compartments and sleep apnea severity in circulating leptin levels: the Icelandic Sleep Apnea Cohort study. *Int J Obesity (Lond)*, **37**: 835-42.
13. Bao R, Huang L, Andrade J, et al. (2014). Review of current methods, applications, and data management for the bioinformatics analysis of whole exome sequencing. *Cancer Inform*, **13(Suppl 2)**: 67-82.
14. Bartholomew D, Knott M, Moustaki I (2011). *Latent variable models and factor analysis.* Wiley: N.J.
15. Basu S, Pan W (2011). Comparison of statistical tests for disease association with rare variants. *Genet Epidemiol*, **35**: 606-19.

16. Benjamini Y, Hochberg Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B*, **57**: 289-300.
17. Berndt J, Klöting N, Kralisch S, et al. (2005). Plasma visfatin concentrations and fat depot-specific mRNA expression in humans. *Diabetes*, **54**: 2911-6.
18. Blangero J, Teslovich TM, Sim X, et al. (2016). Omics-squared: human genomic, transcriptomic and phenotypic data for Genetic Analysis Workshop 19. *BMC Proc*, **10(Suppl 7)**: 71.
19. Boeing H, Wahrendorf J, Becker N (1999). EPIC-Germany - A source for studies into diet and risk of chronic diseases. European Investigation into Cancer and Nutrition. *Ann Nutr Metab*, **43**: 195-204.
20. Bollen KA (1989). *Structural equations with latent variables.* New York: John Wiley & Sons.
21. Bosy-Westphal A, Later W, Hitze B, et al. (2008). Accuracy of bioelectrical impedance consumer devices for measurement of body composition in comparison to whole body magnetic resonance imaging and dual X-ray absorptiometry. *Obes Facts*, **1**: 319-24.
22. Bottolo L, Chadeau-Hyam M, Hastie DI, et al. (2013). GUESS-ing polygenic associations with multiple phenotypes using a GPU-based evolutionary stochastic search algorithm. *PLoS Genet*, **9(8)**: e1003657.
23. Broadaway KA, Cutler DJ, Duncan R (2016). A statistical approach for testing cross-phenotype effects of rare variants. *Am J Hum Genet*, **98(3)**: 525-40.
24. Bredella MA, Ghomi RH, Thomas BJ, et al. (2010). Comparison of DXA and CT in the assessment of body composition in premenopausal women with obesity and anorexia nervosa. *Obesity (Silver Spring)*, **8**: 2227-33.
25. Buckner J, Wilson J, Seligman M, et al. (2010). The gputools package enables GPU computing in R. *Bioinformatics*, **26(1)**: 134-5.
26. Burchard A, Hajaiej H (2006). Rearrangement inequalities for functionals with monotone integrands. *J Funct Anal*, **233(2)**: 561-82.
27. Campbell KL, Foster-Schubert KE, Makar KW, et al. (2013). Gene expression changes in adipose tissue with diet-and/or exercise-induced weight loss. *Cancer Prev Res (Phila)*, **6**: 217-31.
28. Carley H, Taylor MD (2002). A new proof of Sklar's theorem. In: C.M. Cuadras, J. Fortiana, J.A. Rodriguez-Lallena (eds.) *Distributions with given marginals and statistical modelling*, p. 29-34. Kluwer Acad. Publ., Dordrecht.
29. Castro JP, El-Atat FA, McFarlane SI, et al. (2003). Cardiometabolic syndrome: pathophysiology and treatment. *Curr Hypertens Rep*, **5(5)**: 393-401.
30. Chan JL, Blüher S, Yiannakouris N, et al. (2002). Regulation of circulating soluble leptin receptor levels by gender, adiposity, sex steroids, and leptin. *Diabetes*, **51**: 2105-12.
31. Chang W, Luraschi J (2017). *profvis: Interactive Visualizations for Profiling R Code.* R package version 0.3.3. `https://CRAN.R-project.org/package=profvis`.
32. Chen LS, Hsu L, Gamazon ER, et al. (2012). An exponential combination procedure for set-based association tests in sequencing studies. *Am J Hum Genet*, **91**: 977-86.
33. Clayton DG (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, **65**: 141-151.

34. Cole S, Hernán M (2002). Fallibility in estimating direct effects. *Int J Epidemiol*, **31**: 163-5.
35. Conesa A, Madrigal P, Tarazona S, et al. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biol*, **17**: 13.
36. Corradin O, Cohen AJ, Luppino JM, et al. (2016). Modeling disease risk through analysis of physical interactions between genetic variants within chromatin regulatory circuitry. *Nat Genet*, **48**: 1313-20.
37. Dekker JM, Funahashi T, Nijpels G, et al. (2008). Prognostic value of adiponectin for cardiovascular disease and mortality. *J Clin Endocrinol Metab*, **93(4)**: 1489-96.
38. Derkach A, Lawless JF, Sun L (2013). Robust and powerful tests for rare variants using Fisher's method to combine evidence of association from two or more complementary tests. *Genet Epidemiol*, **37**: 110-21.
39. Dummer PD, Limou S, Rosenberg AZ, et al. (2015). APOL1 kidney disease risk variants - an evolving landscape. *Semin Nephrol*, **35(3)**: 222-36.
40. Durinck S, Spellman P, Birney E, et al. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc*, **4**: 1184-91.
41. Efron B (1981). Nonparametric estimates of standard error: the jackknife, the bootstrap, and other methods. *Biometrika*, **68**: 589-99.
42. Ehret GB, Ferreira T, Chasman DI, et al. (2016). The genetics of blood pressure regulation and its target organs from association studies in 342,415 individuals. *Nat Genet*, **48**: 1171-84.
43. Eu-ahsunthornwattana J, Miller EN, Fakiola M, et al. (2014). Comparison of methods to account for relatedness in genome-wide association studies with family-based data. *PLoS Genet*, **10**: e1004445.
44. Everaert C, Luypaert M, Maag JLV, et al. (2017). Benchmarking of RNA-sequencing analysis workflows using whole-transcriptome RT-qPCR expression data. *Sci Rep*, **7**: 1559.
45. Fantuzzi G, Mazzone T (2007). *Adipose tissue and adipokines in health and disease.* Humana Press Inc: New Jersey, US.
46. Fasshauer M, Blüher M (2015). Adipokines in health and disease. *Trends Pharmacol Sci*, **36(7)**: 461-70.
47. Feil R, Fraga MF (2012). Epigenetics and the environment: emerging patterns and implications. *Nat Rev Genet*, **13**: 97-109.
48. Fisman EZ, Tenenbaum A (2014). Adiponectin: a manifold therapeutic target for metabolic syndrome, diabetes, and coronary disease? *Cardiovasc Diabetol*, **13**: 103-12.
49. Fisher M (2006). Cardiometabolic disease: the new challenge? *Pract Diab Int*, **23**: 95-7.
50. Fox CS, Massaro JM, Hoffmann U, et al. (2007). Abdominal visceral and subcutaneous adipose tissue compartments: association with metabolic risk factors in the Framingham Heart Study. *Circulation*, **116**: 39-48.
51. Fox J, Weisberg S (2011). *An R companion to applied regression.* Thousand Oaks CA: Sage.
52. Fredriksson J, Carlsson E, Orho-Melander M, et al. (2006). A polymorphism in the adiponectin gene influences adiponectin expression levels in visceral fat in obese subjects. *Int J Obesity (Lond)*, **30**: 226-32.

53. Friedman J (2016). The long road to leptin. *J Clin Invest*, **126(12)**: 4727-34.
54. Fuchsberger C, Flannick J, Teslovich TM, et al. (2016). The genetic architecture of type 2 diabetes. *Nature*, **536**: 41-7.
55. Furuhashi M, Hotamisligil GS (2008). Fatty acid-binding proteins: role in metabolic diseases and potential as drug targets. *Nat Rev Drug Discov*, **7**: 489-503.
56. Galesloot TE, van Steen K, Kiemeney LA, et al. (2014). A comparison of multivariate genome-wide association methods. *PLoS One*, **9(4)**: e95923.
57. Gene Ontology Consortium (2001). Creating the Gene Ontology resource: design and implementation. *Genome Res*, **11**: 1425-33.
58. Genest C, Rivest LP (1993). Statistical inference procedures for bivariate Archimedean copulas. *J Am Stat Assoc*, **88**: 1034-43.
59. Genovese G, Friedman DJ, Ross MD, et al. (2010). Association of trypanolytic ApoL1 variants with kidney disease in african-americans. *Science*, **329(5993)**: 841-5.
60. Gentle JE (2009). *Computational statistics*. New York: Springer.
61. Goetgeluk S, Vansteelandt S, Goetghebeur E (2008). Estimation of controlled direct effects. *J R Stat Soc Ser B*, **70**: 1049-66.
62. Goldstein BJ, Scalia RG, Ma XL (2009). Protective vascular and myocardial effects of adiponectin. *Nat Clin Pract Cardiovasc Med*, **6**: 27-35.
63. Goldstone AP, Brynes AE, Thomas EL, et al. (2002). Resting metabolic rate, plasma leptin concentrations, leptin receptor expression, and adipose tissue measured by whole-body magnetic resonance imaging in women with Prader-Willi syndrome. *Am J Clin Nutr*, **75**: 468-75.
64. Gomez-Cabrero D, Abugessaisa I, Maier D, et al. (2014). Data integration in the era of omics: current and future challenges. *BMC Syst Biol*, **8(Suppl 2)**: I1.
65. Goodwin S, McPherson JD, McCombie WR (2016). Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet*, **17(6)**: 333-51.
66. Gorlov IP, Gorlova OY, Frazier ML, et al. (2011). Evolutionary evidence of the effect of rare variants on disease etiology. *Clin Genet*, **79**: 199-206.
67. Gottschald M, Knüppel S, Boeing H, et al. (2016). The influence of adjustment for energy misreporting on relations of cake and cookie intake with cardiometabolic disease risk factors. *Eur J Clin Nutr*, **70**: 1318-24.
68. Grossmann S, Bauer S, Robinson PN, et al. (2007). Improved detection of overrepresentation of gene-ontology annotations with parent child analysis. *Bioinformatics*, **23**: 3024.
69. Gumbel EJ (1960). Distribution des valeurs extrêmes en plusieurs dimensions. *Publications de l'Institut de statistique de l'Université de Paris*, **9**: 171-3.
70. Hansen KD, Irizarry RA, Zhijin W (2012). Removing technical variability in RNA-Seq data using conditional quantile normalization. *Biostatistics*, **13(2)**: 204-16.
71. Helgadottir A, Gretarsdottir S, Thorleifsson G, et al. (2016). Variants with large effect on blood lipids and the role of cholesterol and triglycerides in coronary disease. *Nat Genet*, **48**: 634-9.
72. Hindorff LA, MacArthur J, Morales J, et al. (2017). *A Catalog of Published Genome-Wide Association Studies*. Available at: `www.genome.gov/gwastudies`. Accessed 01/06/2017.

73. Hong HR, Ha CD, Jin YY, et al. (2015). The effect of physical activity on serum IL-6 and vaspin levels in late elementary school children. *J Exerc Nutrition Biochem*, **19**: 99-106.
74. Hotamisligil GS, Bernlohr DA (2015). Metabolic functions of FABPs - mechanisms and therapeutic implications. *Nat Rev Endocrinol*, **11**: 592-605.
75. Hotelling H (1933). Analysis of a complex of statistical variables into principal components. *J Educ Psychol*, **24**: 417-41, 498-520.
76. Hotelling H (1936). Relations between two sets of variants. *Biometrika*, **28**: 321-377.
77. Hu CAA, Klopfer EI, Ray PE (2012). Human Apolipoprotein L1 (ApoL1) in cancer and chronic kidney disease. *FEBS Lett*, **586(7)**: 947-55.
78. Huan T, Esko T, Peters MJ, et al. (2015a). A meta-analysis of gene expression signatures of blood pressure and hypertension. *PLoS Genet*, **11(3)**: e1005035.
79. Huan T, Meng Q, Saleh MA, et al. (2015b). Integrative network analysis reveals molecular mechanisms of blood pressure regulation. *Mol Syst Biol*, **11(1)**: 799.
80. Hubbard T, Barker D, Birney E, et al. (2002). The Ensembl genome database project. *Nucleic Acids Res*, **30(1)**: 38-41.
81. Huh JY, Park YJ, Ham M, et al. (2014). Crosstalk between adipocytes and immune cells in adipose tissue inflammation and metabolic dysregulation in obesity. *Mol Cells*, **37(5)**: 365-71.
82. Hunt KA, Mistry V, Bockett NA, et al. (2013). Negligible impact of rare autoimmune-locus coding-region variants on missing heritability. *Nature*, **498**: 232-5.
83. Hunter CA, Kastelein R (2012). Interleukin-27: balancing protective and pathological immunity. *Immunity*, **37**: 960-9.
84. Hunter CA, Jones SA (2015). IL-6 as a keystone cytokine in health and disease. *Nat Immunol*, **16**: 448-57.
85. Hwang S, Kim E, Lee I, et al. (2015). Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Sci Rep*, **5**: 17875.
86. Ibrahim MM (2010). Subcutaneous and visceral adipose tissue: structural and functional differences. *Obes Rev*, **11**: 11-8.
87. Ishimura S, Furuhashi M, Watanabe Y, et al. (2013). Circulating levels of fatty acid-binding protein family and metabolic phenotype in the general population. *PLoS One*, **8**: e81318.
88. Jamaluddin S, Weakley SM, Yao Q, et al. (2012). Resistin: functional roles and therapeutic considerations for cardiovascular disease. *Br J Pharmacol*, **165**: 622-32.
89. Jiang J, Zhang Q, Ma L, et al. (2015). Joint prediction of multiple quantitative traits using a Bayesian multivariate antedependence model. *Heredity (Edinb)*, **115(1)**: 29-36.
90. Joe H (1997). *Multivariate models and multivariate dependence concepts*. Chapman & Hall: London.
91. Jolliffe IT (2002). *Principal component analysis*. Springer: New York.
92. Kaakinen M, Mägi R, Fischer K, et al. (2017). A rare-variant test for high-dimensional data. *Eur J Hum Genet*, **25(8)**: 988-94.
93. Karlsson AK, Kullberg J, Stokland E, et al. (2013). Measurements of total and regional body composition in preschool children: a comparison of MRI, DXA, and anthropometric data. *Obesity (Silver Spring)*, **21**: 1018-24.

94. Kato N, Loh M, Takeuchi F, et al. (2015). Trans-ancestry genome-wide association study identifies 12 genetic loci influencing blood pressure and implicates a role for DNA methylation. *Nat Genet*, **47**: 1282–93.
95. Kearney PM, Whelton M, Reynolds K, et al. (2005). Global burden of hypertension: analysis of worldwide data. *Lancet*, **365(9455)**: 217-23.
96. Kern PA, Ranganathan S, Li C, et al. (2001). Adipose tissue tumor necrosis factor and interleukin-6 expression in human obesity and insulin resistance. *Am J Physiol Endocrinol Metab*, **280(5)**: E745-51.
97. Kim J, Pan W. (2016). *GEEaSPU: Adaptive association tests for multiple phenotypes using generalized estimating equations (GEE)*. R package version 1.0.2. `https://CRAN.R-project.org/package=GEEaSPU`.
98. Kim D, Pertea G, Trapnell C, et al. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*, **14(4)**: R36.
99. Kim J, Zhang Y, Pan W, et al. (2016). Powerful and adaptive testing for multi-trait and multi-SNP associations with GWAS and sequencing data. *Genetics*, **203(2)**: 715-31.
100. Kinnamon DD, Hershberger RE, Martin ER (2012). Reconsidering association testing methods using single-variant test statistics as alternatives to pooling tests for sequence data with rare variants. *PLoS One*, **7(2)**: e30238.
101. Konigorski S, Yilmaz YE, Bull SB (2014). Bivariate genetic association analysis of systolic and diastolic blood pressure by copula models. *BMC Proc*, **8(Suppl 1)**: S72-7.
102. Konigorski S, Yilmaz YE, Pischon T (2016). Genetic association analysis based on a joint model of gene expression and blood pressure. *BMC Proc*, **10(Suppl 7)**: 289-94.
103. Konigorski S, Yilmaz YE, Pischon T (2017). Comparison of single-marker and multi-marker tests in rare variant association studies of quantitative traits. *PLoS One*, **12(5)**: e0178504.
104. Kovacova Z, Tencerova M, Roussel B, et al. (2012). The impact of obesity on secretion of adiponectin multimeric isoforms differs in visceral and subcutaneous adipose tissue. *Int J Obesity (Lond)*, **36**: 1360-5.
105. Kristensen VN, Lingjaerde OC, Russnes HG, et al. (2014). Principles and methods of integrative genomic analyses in cancer. *Nat Rev Cancer*, **14(5)**: 299-313.
106. Kursawe R, Narayan D, Cali AM, et al. (2010). Downregulation of ADIPOQ and PPARγ2 gene expression in subcutaneous adipose tissue of obese adolescents with hepatic steatosis. *Obesity (Silver Spring)*, **18**: 1911-7.
107. Ladouceur M, Dastani Z, Aulchenko YS, et al. (2012). The empirical power of rare variant association methods: results from Sanger sequencing in 1,998 individuals. *PLoS Genet*, **8(2)**: e1002496.
108. Laird NM, Lange C (2011). *The fundamentals of modern statistical genetics*. New York: Springer.
109. Langmead B, Trapnell C, Pop M, et al. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, **10(3)**: R25.
110. Langmead B, Salzberg SL (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods*, **14**: 357-9.
111. Lee S, Kuk JL (2013). Changes in fat and skeletal muscle with exercise training in obese adolescents: comparison of whole-body MRI and dual energy X-ray absorptiometry. *Obesity (Silver Spring)*, **21**: 2063-71.

112. Lee S, Seo CH, Lim B, et al. (2011). Accurate quantification of transcriptome from RNA-Seq data by effective length normalization. *Nucleic Acids Res*, **39(2)**: e9.
113. Lee S, Wu MC, Lin X (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, **13**: 762-75.
114. Lee S, Abecasis GR, Boehnke M, et al. (2014). Rare-variant association analysis: study designs and statistical tests. *Am J Hum Genet*, **95**: 5-23.
115. Lee S, Miropolsky L, Wu MC (2016). *SKAT: SNP-Set (Sequence) Kernel Association Test*. Version 1.2.1 [R package]. Available from: `https://CRAN.R-project.org/package=SKAT`.
116. Liang X, Wang Z Q, Zhang S (2016). An adaptive Fisher's combination method for joint analysis of multiple phenotypes in association studies. *Sci Rep*, **6**: 34323.
117. Li B, Dewey CN (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**: 323.
118. Li B, Leal SM. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet*, **83**: 311-21.
119. Li H, Handsaker B, Wysoker A, et al. (2009). The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, **25**: 2078-9.
120. Li H (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27(21)**: 2987-93.
121. Lihn AS, Bruun JM, He G, et al. (2004). Lower expression of adiponectin mRNA in visceral adipose tissue in lean and obese subjects. *Mol Cell Endocrinol*, **219**: 9-15.
122. Lin DY, Tang ZZ (2011). A general framework for detecting disease associations with rare variants in sequencing studies. *Am J Hum Genet*, **89**: 354-67.
123. Lipman PJ, Lange C (2011). CGene: an R package for implementation of causal genetic analyses. *Eur J Hum Genet*, **19**: 1292-4.
124. Lipman PJ, Liu K, Muehlschlegel JD, et al. (2011). Inferring genetic causal effects on survival data with associated endo-phenotypes. *Genet Epidemiol*, **35**: 119-24.
125. Lippert C, Listgarten J, Liu Y, et al. (2011). FaST linear mixed models for genome-wide association studies. *Nat Methods*, **8**: 833-5.
126. Listgarten J, Lippert C, Kadie CM, et al. (2012). Improved linear mixed models for genome-wide association studies. *Nat Methods*, **9(6)**: 525-6.
127. Liu C, Kraja AT, Smith JA, et al. (2016). Meta-analysis identifies common and rare variants influencing blood pressure and overlapping with metabolic trait loci. *Nat Genet*, **48**: 1162-70.
128. Liu M, Liu F (2012). Up- and down-regulation of adiponectin expression and multimerization: Mechanisms and therapeutic implication. *Biochimie*, **94**: 2126-30.
129. Livak KJ, Schmittgen TD (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) method. *Methods*, **25**: 402-8.
130. Locke AE, Kahali B, Berndt SI, et al. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature*, **518**: 197-206.
131. Lohmueller KE, Sparsø T, Li Q, et al. (2013). Whole-exome sequencing of 2,000 Danish individuals and the role of rare coding variants in type 2 diabetes. *Am J Hum Genet*, **93**: 1072-86.

132. Lotta LA, Scott RA, Sharp SJ, et al. (2016). Genetic predisposition to an impaired metabolism of the branched-chain amino acids and risk of type 2 diabetes: a mendelian randomization analysis. *PLoS Med*, **13**: e1002179.
133. Maahs DM, Ogden LG, Snell-Bergeon JK, et al. (2007). Determinants of serum adiponectin in persons with and without type 1 diabetes. *Am J Epidemiol*, **166**: 731-40.
134. Madsen BE, Browning SR (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet*, **5**: e1000384.
135. Madhavan S, O'Toole JF (2014). The biology of APOL1 with insights into the association of APOL1 variants with chronic kidney disease. *Clin Exp Nephrol*, **18(2)**: 238-42.
136. Magni P, Liuzzi A, Ruscica M, et al. (2005). Free and bound plasma leptin in normal weight and obese men and women: relationship with body composition, resting energy expenditure, insulin-sensitivity, lipid profile and macronutrient preference. *Clin Endocrinol (Oxf)*, **62(2)**: 189-96.
137. Mandt S, Wenzel F, Nakajima S, et al. (2017). Sparse probit linear mixed model. *Machine Learning*, **106(9-10)**: 1621-42.
138. Mantzoros CS, Magkos F, Brinkoetter M, et al. (2011). Leptin in human physiology and pathophysiology. *Am J Physiol Endocrinol Metab*, **301(4)**: E567-84.
139. Mardis ER (2017). DNA sequencing technologies: 2006-2016. *Nat Protoc*, **12**: 213-8.
140. Marioni JC, Mason CE, Mane SM, et al. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*, **18(9)**: 1509-17.
141. Martinussen T, Vansteelandt S, Gerster M, et al. (2011). Estimation of direct effects for survival data by using the Aalen additive hazards model. *J R Stat Soc Ser B*, **73**: 773-88.
142. MacArthur J, Bowler E, Cerezo M, et al. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res*, **45(Database issue)**: D896-901.
143. McCurdy CE, Schenk S, Hetrick B, et al. (2016). Maternal obesity reduces oxidative capacity in fetal skeletal muscle of Japanese macaques. *JCI Insight*, **1(16)**: e86612.
144. Mieth B, Kloft M, Rodríguez JA, et al. (2016). Combining multiple hypothesis testing with machine learning increases the statistical power of genome-wide association studies. *Sci Rep*, **6**: 36671.
145. Morgenthaler S, Thilly WG (2007). A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat Res*, **615**: 28-56.
146. Morris AP, Zeggini E (2010). An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet Epidemiol*, **34**: 188-93.
147. Mortazavi A, Williams BA, McCue K, et al. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*, **5(7)**: 621-628.
148. Nash, JC (2014). On best practice optimization methods in R. *J Stat Softw*, **60(2)**: 1-14.
149. Neale BM, Rivas MA, Voight BF, et al. (2011). Testing for an unusual distribution of rare variants. *PLoS Genet*, **7**: e1001322.

150. Neamat-Allah, Wald D, Huesing A, et al. (2014). Validation of anthropometric indices of adiposity against whole-body magnetic resonance imaging - a study within the German European Prospective Investigation into Cancer and Nutrition (EPIC) cohorts. *PLoS One*, **9**: e91586.
151. Neamat-Allah J, Johnson T, Nabers D, et al. (2015). Can the use of blood-based biomarkers in addition to anthropometric indices substantially improve the prediction of visceral fat volume as measured by magnetic resonance imaging? *Eur J Nutr*, **54(5)**: 701-8.
152. Nelsen RB (2006). *An introduction to copulas*. Springer, New York.
153. Nelson MR, Wegmann D, Ehm MG, et al. (2012). An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science*, **337**: 100-4.
154. Norata GD, Ongari M, Garlaschelli K, et al. (2007). Plasma resistin levels correlate with determinants of the metabolic syndrome. *Eur J Endocrinol*, **156**: 279-84.
155. O'Brien PC (1984). Procedures for comparing samples with multiple endpoints. *Biometrics*, **40(4)**: 1079-87.
156. O'Reilly PF, Hoggart CJ, Pomyen Y, et al. (2012). Multi-Phen: joint model of multiple phenotypes can increase discovery in GWAS. *PLoS One*, **7**: e34861.
157. Ott J, Rabinowitz D (1999). A principal-components approach based on heritability for combining phenotype information. *Hum Hered*, **49(2)**: 106-11.
158. Ouchi N, Parker JL, Lugus JJ, et al. (2011). Adipokines in inflammation and metabolic disease. *Nat Rev Immunol*, **11**: 85-91.
159. Ozsolak F, Milos PM (2010). RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet*, **12(2)**: 87-98.
160. Pai JK, Pischon T, Ma J, et al. (2004). Inflammatory markers and the risk of coronary heart disease in men and women. *N Engl J Med*, **351**: 2599-610.
161. Pan W (2009). Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Genet Epidemiol*, **33(6)**: 497-507.
162. Pan W, Kwak IY, Wei P (2015). A powerful pathway-based adaptive test for genetic association with common or rare variants. *Am J Hum Genet*, **97(1)**: 86-98.
163. Pan W, Kim J, Zhang Y, et al. (2014). A powerful and adaptive association test for rare variants. *Genetics*, **197(4)**: 1081-95.
164. Pearl J (1995). Causal diagrams for empirical research. *Biometrika*, **82**: 669-88.
165. Pearson K (1901). On lines and planes of closest fit to systems of points in space. *Phil Mag*, **6(2)**: 559-72.
166. Pickrell JP, Berisa T, Liu JZ, et al. (2016). Detection and interpretation of shared genetic influences on 42 human traits. *Nat Genet*, **48**: 709-17.
167. Pietiläinen KH, Kannisto K, Korsheninnikova E, et al. (2006). Acquired obesity increases CD68 and tumor necrosis factor-alpha and decreases adiponectin gene expression in adipose tissue: a study in monozygotic twins. *J Clin Endocrinol Metab*, **19**: 2776-81.
168. Pirooznia M, Kramer M, Parla J, et al. (2014). Validation and assessment of variant calling pipelines for next-generation sequencing. *Hum Genomics*, **8**: 14.
169. Pischon T (2009). Use of obesity biomarkers in cardiovascular epidemiology. *Dis Markers*, **26**: 247-63.
170. Pischon T, Bamberger CM, Kratzsch J, et al. (2005). Association of plasma resistin levels with coronary heart disease in women. *Obes Res*, **13(10)**: 1764-71.

171. Pischon T, Girman CJ, Hotamisligil GS, et al. (2004). Plasma adiponectin levels and risk of myocardial infarction in men. *JAMA*, **291(14)**: 1730-7.
172. Price AL, Kryukov GV, de Bakker PIW, et al. (2010). Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet*, **86**: 832-8.
173. Price AL, Patterson NJ, Plenge RM, et al. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*, **38**: 904-9.
174. Purcell SM, Moran JL, Fromer M, et al. (2014). A polygenic burden of rare disruptive mutations in schizophrenia. *Nature*, **506**: 185-90.
175. R Core Team (2017). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.
176. Relton CL, Davey Smith G (2012a). Is epidemiology ready for epigenetics? *Int J Epidemiol*, **41**: 5-9.
177. Relton CL, Davey Smith G (2012b). Two-step epigenetic Mendelian randomization: a strategy for establishing the causal role of epigenetic processes in pathways to disease. *Int J Epidemiol*, **41**: 161-76.
178. Riboli E, Kaaks R (1997). The EPIC Project: rationale and study design. European Prospective Investigation into Cancer and Nutrition. *Int J Epidemiol*, **26(Suppl 1)**: S6-14.
179. Ried JS, Jeff MJ, Chu AY, et al. (2016). A principal component meta-analysis on multiple anthropometric traits identifies novel loci for body shape. *Nat Commun*, **7**: 13357.
180. Risso D, Schwartz K, Sherlock G, et al. (2011). GC-content normalization for RNA-Seq data. *BMC Bioinformatics*, **12**: 480.
181. Ritchie MD, Holzinger ER, Li R, et al. (2015). Methods of integrating data to uncover genotype-phenotype interactions. *Nat Rev Genet*, **16(2)**: 85-97.
182. Ritz C, Streibig J (2005). Bioassay analysis using R. *J Stat Softw*, **12**: 1-22.
183. Robins JM (1986). A new approach to causal inference in mortality studies with a sustained exposure period - Application to control of the healthy worker survivor effect. Mathematical models in medicine: Diseases and epidemics. Part 2. *Math Modelling*, **7**: 1393-512.
184. Robins JM (1992). Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. *Biometrika*, **79**: 321-34.
185. Robins JM, Greenland S (1994). Adjusting for differential rates of PCP prophylaxis in high- versus low-dose. AZT treatment arms in an AIDS randomized trial. *J Am Statist Ass*, **89**: 737-49.
186. Robins JM, Hernán MA, Brumback B (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, **11**: 550-60.
187. Robinson MD, Oshlack A (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*, **11(3)**: R25.
188. Rosenbaum P (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *J R Stat Soc Ser A*, **147**: 656-66.
189. Ross R (2003). Advances in the application of imaging methods in applied and clinical physiology. *Acta Diabetol*, **40(Suppl 1)**: S45-50.

190. Rosseel Y (2012). lavaan: an R package for structural equation modeling. *J Stat Softs*, **48(2)**: 1-36.
191. Rüschendorf L (2009). On the distributional transform, Sklar's Theorem, and the empirical copula process. *J Statist Plan Infer*, **139(11)**: 39217.
192. Savage DB, Sewter CP, Klenk ES, et al. (2001). Resistin / Fizz3 expression in relation to obesity and peroxisome proliferator-activated receptor-gamma action in humans. *Diabetes*, **50(10)**: 2199-202.
193. Schillert A, Konigorski S (2016). Joint analysis of multiple phenotypes - summary of results and discussions from the Genetic Analysis Workshop 19. *BMC Genet*, **17(Suppl 2)**: 7.
194. Schweizer B, Sklar A (1974). Operations on distribution functions not derivable from operations on random variables. *Studia Math*, **52**: 43-52.
195. Self GS, Liang KY (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J Am Stat Assoc*, **82**: 605-10.
196. Shih JH, Louis TA (1995). Inferences on the association parameter in copula models for bivariate survival data. *Biometrics*, **51**: 1384-99.
197. Shriner D (2012). Moving toward system genetics through multiple trait analysis in genome-wide association studies. *Front Genet*, **3**: 1.
198. Shungin D, Winkler TW, Croteau-Chonka DC, et al. (2015). New genetic loci link adipose and insulin biology to body fat distribution. *Nature*, **518**: 187-196.
199. Sklar A (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de statistique de l'Université de Paris 8*: 229-31.
200. Smith GD, Ebrahim S (2003). 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol*, **32**: 1-22.
201. Spranger J, Kroke A, Möhlig M, et al. (2003). Adiponectin and protection against type 2 diabetes mellitus. *Lancet*, **361(9353)**: 226-8.
202. Steppan CM, Bailey ST, Bhat S, et al. (2001). The hormone resistin links obesity to diabetes. *Nature*, **409(6818)**: 307-12.
203. Suchard MA, Wang Q, Chan C, et al. (2010). Understanding GPU programming for statistical computation: studies in massively parallel massive mixtures. *J Comp Graph Stat*, **19(2)**: 419-38.
204. Sun J, Bhatnagar S, Oualkacha K, et al. (2016a). Joint analysis of multiple blood pressure phenotypes in GAW19 data by using a multivariate rare-variant association test. *BMC Proc*, **10(Suppl 7)**: 309-13.
205. Sun J, Oualkacha K, Forgetta V, et al. (2016b). A method for analyzing multiple continuous phenotypes in rare variant association studies allowing for flexible correlations in variant effects. *Eur J Hum Genet*, **24(9)**: 1344-51.
206. Sun J, Zheng Y, Hsu L (2013). A unified mixed-effects model for rare-variant association in sequencing studies. *Genet Epidemiol*, **37**: 334-44.
207. Sun L, Dimitromanolakis A, Faye LL, et al. (2011). BR-Squared: a practical solution to the winner's curse in genome-wide scans. *Hum Genet*, **129**: 545-52.
208. Sun Q, van Dam RM, Meigs JB, et al. (2009). Leptin and soluble leptin receptor levels in plasma and risk of type 2 diabetes in U.S. women: a prospective study. *Diabetes*, **59(3)**: 611-8.

209. Surendran P, Drenos F, Young R, et al. (2016). Trans-ancestry meta-analyses identify rare and common variants associated with blood pressure and hypertension. *Nat Genet*, **48**: 1151-61.
210. Sutinen J, Korsheninnikova E, Funahashi T, et al. (2003). Circulating concentration of adiponectin and its expression in subcutaneous adipose tissue in patients with highly active antiretroviral therapy-associated lipodystrophy. *J Clin Endocrinol Metab*, **88**: 1907-10.
211. Szymczak S, Biernacka JM, Cordell HJ, et al. (2009). Machine learning in genome-wide association studies. *Genet Epidemiol*, **33(Suppl 1)**: S51-7.
212. Tam CS, Lecoultre V, Ravussin E (2011). Novel strategy for the use of leptin for obesity therapy. *Expert Opin Biol Ther*, **11(12)**: 1677-85.
213. Taylor AE, Kuper H, Varma RD, et al. (2012). Validation of dual energy X-ray absorptiometry measures of abdominal fat by comparison with magnetic resonance imaging in an Indian population. *PLoS One*, **7**: e51042.
214. Terra X, Quintero Y, Auguet T, et al. (2011). FABP 4 is associated with inflammatory markers and metabolic syndrome in morbidly obese women. *Eur J Endocrinol*, **164**: 539-47.
215. Thomas T, Burguera B, Melton LJ 3rd, et al. (2000). Relationship with serum leptin levels with body composition and sex steroid and insulin levels in men and women. *Metabolism*, **49**: 1278-84.
216. Tobin MD, Sheehan NA, Scurrah KJ, et al. (2005). Adjusting for treatment effects in studies of quantitative traits: antihypertensive therapy and systolic blood pressure. *Stat Med*, **24**: 2911-35.
217. Trapnell C, Pachter L, Salzberg SL (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25(9)**: 1105-11.
218. Tsai WC, Wu HY, Peng YS, et al. (2017). Association of intensive blood pressure control and kidney disease progression in nondiabetic patients with chronic kidney disease: a systematic review and meta-analysis. *JAMA Intern Med*, **177(6)**: 792-9.
219. UK10K Consortium (2015). The UK10K project identifies rare variants in health and disease. *Nature*, **526**: 82-90.
220. van der Sluis S, Posthuma D, Dolan CV (2013). TATES: efficient multivariate genotype-phenotype analysis for genome-wide association studies. *PLoS Genet*, **9(1)**: e1003235.
221. Vansteelandt S, Goetgeluk S, Lutz S, et al. (2009). On the adjustment for covariates in genetic association analysis: a novel, simple principle to infer direct causal effects. *Genet Epidemiol*, **33**: 394-405.
222. Vansteelandt S, Joffe M (2014). Structural nested models and G-estimation: the partially realized promise. *Stat Sci*, **29**: 707-31.
223. Vitak SA, Torkenczy KA, Rosenkrantz JL, et al. (2017). Sequencing thousands of single-cell genomes with combinatorial indexing. *Nat Methods*, **14**: 302-8.
224. Wagner GP, Kin K, Lynch VJ (2012). Measurement of mRNA abundance using RNA-Seq data: RPKM measure is inconsistent among samples. *Theory Biosci*, **131(4)**: 281-5.
225. Wald D, Teucher B, Dinkel J, et al. (2012). Automatic quantification of subcutaneous and visceral adipose tissue from whole-body magnetic resonance images suitable for large cohort studies. *J Magn Reson Imaging*, **36**: 1421-34.

226. Wallenius V, Wallenius K, Ahrén B, et al. (2002). Interleukin-6-deficient mice develop mature-onset obesity. *Nat Med*, **8(1)**: 75-9.
227. Wang Z, Gerstein M, Snyder M (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, **10(1)**: 57-63.
228. Wang Z, Wang X, Sha Q, et al. (2016). Joint analysis of multiple traits in rare variant association studies. *Ann Hum Genet*, **80**: 162-71.
229. Wannamethee SG, Whincup PH, Lennon L, et al. (2007). Circulating adiponectin levels and mortality in elderly men with and without cardiovascular disease and heart failure. *Arch Intern Med*, **167(14)**: 1510-7.
230. Warnes G, Gorjanc G, Leisch F, et al. (2013). *genetics: Population Genetics*. R package version 1.3.8.1.
231. Warren H, Evangelou E, Cabrara CP, et al. (2017). Genome-wide association analysis identifies novel blood pressure loci and offers biological insights into cardiovascular risk. *Nat Genet*, **49**: 403-15.
232. White H (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, **50(1)**: 1-25.
233. Weikert C, Westphal S, Berger K, et al. (2008). Plasma resistin levels and risk of myocardial infarction and ischemic stroke. *J Clin Endocrinol Metab*, **93(7)**: 2647-53.
234. Welter D, MacArthur J, Morales J, et al. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res*, **42**: D1001-6.
235. Wickham H, Chang W (2017). *devtools: tools to make developing R packages easier*. R package version 1.13.2. `https://CRAN.R-project.org/package=devtools`.
236. Wickham H, Danenberg P, Eugster M (2017). *roxygen2: in-line documentation for R*. R package version 6.0.1. `https://CRAN.R-project.org/package=roxygen2`.
237. Wientzek A, Vigl M, Steindorf K, et al. (2014). The improved physical activity index for measuring physical activity in EPIC Germany. *PLoS One*, **9**: e92005.
238. Won JC, Park CY, Lee WY, et al. (2009). Association of plasma levels of resistin with subcutaneous fat mass and markers of inflammation but not with metabolic determinants or insulin resistance. *J Korean Med Sci*, **24**: 695-700.
239. Woodward L, Akoumianakis I, Antoniades C (2016). Unravelling the adiponectin paradox: novel roles of adiponectin in the regulation of cardiovascular disease. *Br J Pharmacol*, doi:10.1111/bph.13619.
240. Wu MC, Lee S, Cai T, et al. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet*, **89**: 82-93.
241. Xie Y (2017). *knitr: a general-purpose package for dynamic report generation in R*. R package version 1.16.
242. Xing G, Lin CY, Wooding SP, et al. (2012). Blindly using Wald's test can miss rare disease-causal variants in case-control association studies. *Ann Hum Genet*, **76**: 168-77.
243. Xu C, Ladouceur M, Dastani Z, et al. (2012). Multiple regression methods show great potential for rare variant association tests. *PLoS One*, **7**: e41694.
244. Xu C, Ciampi A, Greenwood CMT, et al. (2014). Exploring the potential benefits of stratified false discovery rates for region-based testing of association with rare genetic variation. *Front Genet*, **5**: 1-13.
245. Xu H, Barnes GT, Yang Q, et al. (2003a). Chronic inflammation in fat plays a crucial role in the development of obesity-related insulin resistance. *J Clin Invest*, **112(12)**: 1821-30.

246. Xu X, Tian L, Wei LJ (2003b). Combining dependent tests for linkage or association across multiple phenotypic traits. *Biostatistics*, **4(2)**: 223-9.
247. Yamagishi S, Adachi H, Matsui T, et al. (2009). Decreased high-density lipoprotein cholesterol level is an independent correlate of circulating tumor necrosis factor-alpha in a general population. *Clin Cardiol*, **32**: E29-32.
248. Yang G, Fan W, Luo B, et al. (2016). Circulating resistin levels and risk of colorectal cancer: A meta-analysis. *Biomed Res Int*, **2016**: 7367485.
249. Yang IS, Kim S (2015). Analysis of whole transcriptome sequencing data: workflow and software. *Genomics Inform*, **13(4)**: 119-25.
250. Yang Q, Wang Y (2012). Methods for analyzing multivariate phenotypes in genetic association studies. *J Probab Stat*, **2012**: 652569.
251. Yeo IK, Johnson R (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, **87**: 954-9.
252. Zastrow O, Seidel B, Kiess W, et al. (2003). The soluble leptin receptor is crucial for leptin action: evidence from clinical and experimental data. *Int J Obesity (Lond)*, **27**: 1472-8.
253. Zeng CP, Chen YC, Lin X, et al. (2017). Increased identification of novel variants in type 2 diabetes, birth weight and their pleiotropic loci. *J Diabetes*, doi: 10.1111/1753-0407.12510.
254. Zhang J, Scarpace PJ (2009). The soluble leptin receptor neutralizes leptin-mediated STAT3 signalling and anorexic responses in vivo. *Br J Pharmacol*, **158(2)**: 475-82.
255. Zhang JT (2005). Approximate and asymptotic distributions of chi-squared-type mixtures with applications. *J Am Stat Assoc*, **100(469)**: 273-85.
256. Zhang Y, Proenca R, Maffei M, et al. (1994). Positional cloning of the mouse obese gene and its human homologue. *Nature*, **372(6505)**: 425-32.
257. Zhang Y, Xu Z, Shen X, et al. (2014). Testing for association with multiple traits in generalized estimation equations, with application to neuroimaging data. *Neuroimage*, **96**: 309-25.
258. Zhu H, Zhang S, Sha Q (2015). Power comparisons of methods for joint association analysis of multiple phenotypes. *Hum Hered*, **80(3)**: 144-52.

Appendix A

A.1 Selbständigkeitserklärung

Ich erkläre, dass ich die Dissertation selbständig und nur unter Verwendung der von mir gemäß § 7 Abs. 3 der Promotionsordnung der Mathematisch-Naturwissenschaftlichen Fakultät, veröffentlicht im Amtlichen Mitteilungsblatt der Humboldt-Universität zu Berlin Nr. 126/2014 am 18.11.2014 angegebenen Hilfsmittel angefertigt habe.

....
(Date) (Signature)

A.2 Publication list and contributions

1. Schillert A*, Konigorski S* (2016). Joint analysis of multiple phenotypes - summary of results and discussions from the Genetic Analysis Workshop 19. *BMC Genetics*, **17(Suppl 2)**: 7. (*contributed equally)

2. Konigorski S, Yilmaz YE, Pischon T (2016). Genetic association analysis based on a joint model of gene expression and blood pressure. *BMC Proceedings*, **10(Suppl 7)**: 57.

3. Konigorski S, Wang Y, Cigsar C, Yilmaz YE (2016). Estimating and testing direct genetic effects in directed acyclic graphs with multiple phenotypes using estimating equations [Abstract]. In: The 2016 Annual Meeting of the International Genetic Epidemiology Society. *Genetic Epidemiology*, **40**: 609-674.

4. Konigorski S, Yilmaz YE, Pischon T (2017). Comparison of single-marker and multi-marker tests in rare variant association studies of quantitative traits. *PLoS One*, **12(5)**: e0178504.

5. Konigorski S, Wang Y, Cigsar C, Yilmaz YE (2018). Estimating and testing direct genetic effects in directed acyclic graphs using estimating equations. *Genetic Epidemiology*, **42**: 174-186.

6. Konigorski S, Yilmaz YE (2018). Copula-based joint analysis of multiple phenotypes. *R package version 0.1.0.*

7. Konigorski S, Yilmaz YE (2018). CIEE: Estimating and testing direct effects in directed acyclic graphs using estimating equations. *R package version 0.1.1.*

8. Konigorski S, Yilmaz YE, Pischon T (2017). C-JAMP: Copula-based joint analysis of multiple traits of a phenotype to improve the power of rare-variant tests. *Manuscript for publication.*

I hereby confirm the following contribution in these studies:

- Publication 1: AS and SK drafted and wrote the manuscript jointly. Both read and approved the final manuscript.

- Publication 2: SK, YEY and TP designed the overall study; SK conducted statistical analyses; SK and YEY drafted the manuscript. All authors read and approved the final manuscript.

- Publications 3 & 5: SK, YEY and CC designed the overall study; SK conducted all statistical analyses and statistical derivations supervised by YEY, except for the part described in section A.4.3 which was conducted by YEY, CC and YW. SK drafted the manuscript. All authors read and approved the final manuscript.

- Publication 4: Study conceptualization SK YEY TP; writing of all R scripts SK; data curation SK; formal analysis SK; methodology SK YEY; software SK YEY; supervision YEY TP; visualization SK; writing (original draft preparation) SK; writing (review and editing) SK YEY TP.

- Publication 6: SK conceptualized the R package, wrote all R functions, the vignette and documentation, and created the R package. The first draft of the copula likelihood function was based on a function written by YEY. YEY supervised the first draft of the R functions.

- Publication 7: SK conceptualized the R package, wrote all R functions, the vignette, documentation, and created the R package; YEY supervised the first draft of the R functions.

- Publication 8: SK, YEY and TP designed the overall study; SK conducted statistical analyses and SK drafted the manuscript. All authors read and approved the final manuscript.

A.3 Thesis sections in publications

Some of the publication described in the previous section A.2 underlie sections in this thesis:

- Section 2.2.1: The formalization of the tested hypotheses of SMTs and MMTs, and the overview of rare variant tests are decribed in manuscript (4).
- Section 2.2.2.2: The motivation for investigating direct genetic effects and overview about different statistical approaches can be found in manuscript (5).
- Section 3.1: The theoretical details of C-JAMP are partly described in manuscript (2), and in more detail together with the implementation in manuscripts (6), (8).
- Section 3.2: The theoretical details of CIEE are described in manuscript (5), and the implementation in manuscript (7).
- Section 4.1: The simulation study comparing SMTs and MMTs in this section as well as in the supplementary results (sections A.4.1, A.5.2 and A.6.2) are described in manuscript (4).
- Section 4.2: The simulation study comparing C-JAMP with SMTs and MMTs in this section as well as in the supplementary results (sections A.4.2, A.5.3 and A.6.3) are described in manuscript (8).
- Section 4.3: The simulation study comparing CIEE with alternative approaches in this section as well as in the supplementary results (sections A.5.4 and A.6.4) are described in manuscript (5).
- Section 5.1: The application of C-JAMP to the GAW19 data is described in manuscript (2), and the application of CIEE to the GAW19 data is described in manuscript (5).
- Section A.4.3 is described in manuscript (5), has been drafted by YEY, CC and YW, and is reproduced here for completeness.

A.4 Supplementary results and study descriptions

A.4.1 Simulation study: single-marker versus multi-marker tests

A.4.1.1 Power for identifying a causal gene with the SMT and MMTs

For an illustration and explanation of the performance of the SMT described in section 4.1.2.3, I provide more details for the main results of scenario 4 in the following. Here, the SMT had a power of 26.3% to identify a causal gene in a genome-wide study under the nominal α level of 2.5×10^{-6}. That means that within 2,629 replicates (26.3% of the 10,000 replicates), there was at least one rare SNV in a replicate (gene) with adjusted p-value smaller than 2.5×10^{-6}, which led to the identification of the gene as causal. In fact, 3,084 SNVs had adjusted p-value smaller than 2.5×10^{-6} in the 2,629 replicates: 20 (of 15,108) causal singletons were correctly identified, 14 (of 4,675) causal doubletons were correctly identified, 341 (of 5,874) causal SNVs with MAF between 0.001 and 0.005, 428 (of 1,550) causal SNVs with MAF between 0.005 and 0.01, and 2,281 (of 2,989) causal SNVs with MAF between 0.01 and 0.03. If singletons and doubletons had been excluded from the analysis of linear regression, the power would have been slightly elevated to 28.0% in this scenario. On the other hand, for the candidate-gene approach using a significance threshold of 0.05, the gene-level power of the SMT was 73.6% and substantially more very rare variants were identified and drove the high gene-level power. In more detail, 2,576 (of 15,108) causal singletons were correctly identified, 1,551 (of 4,675) causal doubletons were correctly identified, 3,757 (of 5,874) causal SNVs with MAF between 0.001 and 0.005, 1,482 (of 1,550) causal SNVs with MAF between 0.005 and 0.01, and 2,980 (of 2,989) causal SNVs with MAF between 0.01 and 0.03. For this analysis, the power would have been substantially decreased to 62.4% by excluding singletons and doubletons.

We conducted several sensitivity checks to add further details to the comparison. First, phenotype data was generated for scenarios 1-12 under the alternative hypothesis using the same parameter values as in Table 4.1, but for larger sample sizes and based on larger genes. More specifically, I considered the same genetic dataset and the same percentage of causal SNVs but with on average 58 SNVs or 572 SNVs per gene with a sample size of $n = 5,000$. Also, another genetic dataset ('SKAT.example' data in the `SKAT` R package) including 67 SNVs per gene was used to generate data under scenarios 1, 4, 7, 8, 12 with a sample size of $n = 2,000$. The power estimates are shown in Supplementary Tables A.4-A.6, respectively, and increased naturally for all methods. When a gene of average size 33 SNVs with a sample size of $n = 1,000$ was analyzed in comparison to a gene with average size 58 SNVs with a sample size of $n = 5,000$, the increased sample size helped SMTs with a higher power increase compared to MMTs. For example, while the power of the SMT versus SKAT-O was 19.2% versus 19.9% (for $\alpha = 0.05$) and 0.5% versus 0.7% (for $\alpha = 2.5 \times 10^{-6}$) in scenario 2 with $n = 1,000$ (Supplementary Tables A.2-A.3), it was 19.5% versus 15.3% ($\alpha = 2.5 \times 10^{-6}$) for the larger sample and gene size. Hence, for larger sample sizes, SMTs are more powerful than all MMTs also for moderate effect sizes. These conclusions did not change when the same larger sample size ($n = 5,000$) was analyzed with much larger genes including on average 572 SNVs (Supplementary Table A.5). The results of analyzing a sample size of $n = 2,000$ and a gene size of 67 yielded similar conclusions compared to the main analysis including $n = 1,000$ individuals.

As a second sensitivity check, I conducted the analysis under the alternative scenarios by only analyzing rare SNVs (with MAF$\leq$ 0.03) in a gene and disregarding the (non-causal) common SNVs. The power of all approaches increased, but the conclusion which approach has the highest power in each scenario did not change, for both nominal levels (see Supplementary Tables A.2-A.3). As a notable difference, the power of the burden test was closer to the power of SKAT and SKAT-O, which is in line with the results reported in Lee et al. (2012). Furthermore, the power increase of SKAT and SKAT-O was marginally higher than of the SMT. Third, if the more conservative Bonferroni correction is used to account for testing multiple SNVs in a gene with SMTs, the power decreases slightly, but does not change any of the conclusions regarding which method has the highest power in which scenario (see Supplementary Table A.7).

In addition, the same simulation set-up was used to generate a case-control trait from a logistic regression model to compare the SMTs and MMTs for an analysis of a binary trait. Further details are given in the appendix in section A.4.1.1, and results of the power comparison in Supplementary Table A.8. The results indicate that the power of all tests was much lower compared to the analysis of quantitative traits. The decrease was much stronger for the SMT, which had a very low power across all scenarios in line with previous studies (Basu & Pan, 2011; Pan et al., 2014; Xing et al., 2012).

A.4.1.2 Power for identifying a causal variant versus a causal gene

After comparing the single-marker and multi-marker approaches in their ability to identify a causal gene, another important question for practical applications is: What is the power of the tests for identifying a causal variant for a quantitative trait, and how does it compare to the power for identifying a gene containing causal SNVs (i.e., causal gene)? Importantly, this question should not be confused with a power comparison of SMTs (which can be used to identify both causal genes and causal variants) and MMTs (which can be used to identify causal genes). Obtaining SNV-level power estimates for SMTs is straightforward, and Supplementary Table A.9 shows power estimates of SMTs for identifying a causal single variant with a given MAF in comparison to the power for identifying a causal gene (as shown in Figure 4.2). For the interpretation of these results, it should be noted that the power estimates are based on testing different hypotheses and corrected for multiple testing of 20,000 genes in the gene-level analysis and 500,000 SNVs in the variant-level analysis. For variants with low MAF (e.g., MAF$<$ 0.005), the power for identifying a single variant is very low relative to gene-level tests across all scenarios, which corresponds to the reported low power of SMTs in the literature. For SNVs with MAF$\geq$ 0.005, the power of variant-level tests is higher than the power of gene-level tests when there are 5% causal SNVs in a gene, and it is again lower compared to gene-level tests when there are higher percentages of causal SNVs. More specifically, when a variant was tested with 10 observed minor alleles (which equals MAF$=$ 0.005 in this study), then the power for identifying the causal SNV increased and was around 20% for larger effect sizes (scenarios 13, 16, 19, 22 in Supplementary Table A.9) and less than 6% for medium and small effect sizes in the other scenarios.

If SMTs or MMTs are first used to identify genes that are associated with the trait, then SMTs have to be used in a second stage to identify causal SNVs, and the power estimates

should be adjusted to account for the follow-up testing. Developing and evaluating the power of such a 2-stage method was not the focus of this study, and as an upper limit to this power, conditional power estimates are shown in Supplementary Table A.10, obtained by applying the SMT in the second stage to test all SNVs in a gene which has been identified as causal in the first stage. The conditional power was on average 13% for singletons and 85% for SNVs with 10 observed minor alleles for larger effect sizes (scenarios 13, 16, 19, 22 in Supplementary Table A.10), on average 2% and 23% for medium effect sizes (scenarios 14, 17, 20, 23 in Supplementary Table A.10) and on average 1% and 8% (scenarios 15, 18, 21, 24 in Supplementary Table A.10) for smaller effect sizes.

A.4.1.3 Analysis of binary traits

A.4.1.3.1 Data generation For a comparison of the power in studies of quantitative traits and binary traits, case-control phenotypes were generated for the scenarios 1, 4, 7, 8, 12 in Table 4.1 using a logistic regression model, with the same genetic data (SNVs X_j), same covariates Z_1, Z_2, the same effect sizes β_j, and the same percentages of causal variants and direction of effects as for quantitative traits. The disease status Y was generated from a logistic regression model with

$$Y|z_1, z_2, x_1, \ldots, x_k \sim \text{Bernoulli}(p(z_1, z_2, x_1, \ldots, x_k)),$$

where $p(z_1, z_2, x_1, \ldots, x_k) = \frac{\exp(0.5z_1+0.5z_2+\sum \beta_j x_j)}{(1+\exp(0.5z_1+0.5z_2+\sum \beta_j x_j)}$ for the k SNVs X_j in a gene, $Z_1 \sim$ Bin(0.5) which was centered for the data generation, $Z_2 \sim N(0,1)$ and $\beta_j = c \cdot |\log_{10}(\text{MAF}_j)|$.

A.4.1.3.2 Methods To evaluate the performance of SMTs, the following logistic regression model of Y was fitted separately for each SNV X_j in a gene,

$$Pr(Y = 1|z_1, z_2, x_j) = \frac{\exp(\beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_{XY,j} x_j)}{1 + \exp(\beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_{XY,j} x_j)} \tag{A.4.1}$$

using the `glm()` function in R (with default settings), the MLE $\widehat{\beta}_{XY}$ and its standard error estimate $\widehat{SE}(\widehat{\beta}_{XY})$ were obtained and the Wald test was computed for testing $H_0 : \beta_{XY} = 0$ for all SNVs X_j. Adjustments for multiple testing of all SNVs in a gene with the SMT were done using the BH correction and the minimum p-value in a gene was extracted for the gene-level evaluation.

For the burden test, the `glm()` function in R was similarly used with default settings to obtain MLEs and to compute Wald tests for testing $H_0 : \alpha = 0$, by fitting the logistic regression model

$$Pr(Y = 1|z_1, z_2, x_1, \ldots, x_k) = \frac{\exp(\beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_{XY} \sum x_j)}{1 + \exp(\beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_{XY} \sum x_j)}.$$

For SKAT and SKAT-O, the `SKATBinary()` function in the R `SKAT` package was used with default settings.

A.4.1.3.3 Results Regarding the single-marker approach, while ML estimation in the logistic regression model (A.4.1) did not provide valid point estimates and SE estimates for testing singletons and doubletons, the type I errors of the SMT gene-level tests were not inflated and rather very conservative (data not shown). The results of the power comparison

are shown in Supplementary Table A.8 and indicate that the power of the SMT and MMTs was much lower compared to the analysis of quantitative traits. The decrease was much stronger for the SMT, which had a very low power across all scenarios.

A.4.1.4 Analysis of GAW19 data

For a comparison of the SMT and MMTs in a real data application, I performed a gene-level association analysis of systolic blood pressure (SBP) in the Genetic Analysis Workshop 19 data described in section 5.1.1.1. I tested the association of 9 previously identified blood pressure (SBP, diastolic BP, or pulse pressure) genes on chromosome 19 (INSR, RRAS, ZNF101, ELAVL3, RGL3, AMH, DOT1L, PLEKHJ1, SF3A2) (Kato et al., 2015; Ehret et al., 2016; Liu et al., 2016; Surendran et al., 2016) with SBP, using the complete data available for 103 unrelated individuals. Adjusted phenotypes SBP_{adj} were obtained for the analysis as described in section 5.1.1.2.

A query to the Ensembl database with reference genome GRCh37.p13 through BioMart was used to define gene regions as 5kb around the gene start and end positions for the gene-level testing. After basic standard quality checks and exclusion of SNVs with more than 5% missing base calls, this yielded 2,779 common (MAF>0.03) and rare (MAF≤ 0.03) SNVs in the 9 defined gene regions for the gene-level association tests with SBPadj SBP_{adj}. The SMT, SKAT, SKAT-O, and burden test were computed as described in section 4.1: For SKAT and SKAT-O, the `SKAT()` function in the `SKAT` R package was used with default settings. For the SMT, a Wald-type t-test was computed under the linear regression model for each SNV, the p-values in a gene were adjusted with the BH-correction, and the minimum adjusted p-value was extracted for each gene for the gene-level association test. First, all 2,779 common and rare SNVs were included leading to gene sizes between 59 and 1,395 SNVs with an average gene size of 309 SNVs. Second, the analysis was restricted to the 1,746 rare SNVs, which led to an average gene size of 194 SNVs.

In the genetic association analysis of the adjusted SBP phenotype, gene-level hypothesis tests were performed using the SMT, SKAT, SKAT-O, and the burden test. The normal distribution assumption for the error terms in the linear regression of the adjusted SBP phenotype was satisfied. The p-values obtained under each test to assess the significance of the 9 candidate genes are shown in Supplementary Table A.11, for the analyses based on two different inclusion criteria of SNVs: all common and rare SNVs in each gene, and only rare SNVs. The comparison showed that the SMT yielded the smallest p-value for most candidate genes, and was the only approach with a p-value less than 0.05. Furthermore, the power estimates of the SMT, SKAT, and SKAT-O are relatively consistent using either inclusion criterion for SNVs. The burden test, on the other hand, is more strongly influenced by the inclusion criterion. When the analysis was extended to test all 2,858 gene regions on chromosome 19 including 273,636 common and rare SNVs and 218,198 rare SNVs, the same above conclusions were reached but none of the approaches yielded a significant gene-level test for a nominal level of 0.05 corrected for multiple testing.

A.4.2 Simulation study: C-JAMP

A.4.2.1 Finite sample distribution of the C-JAMP estimates and Wald-type test statistic

In this section, the empirical adjustment for the Wald-type test statistic obtained through C-JAMP are described. After obtaining ML parameter estimates through C-JAMP, different test statistics can be used for testing a SNV effect. First investigations showed that using the Wald-type test statistic led to smaller empirical type I errors compared to LRTs (data not shown), so that the Wald-type test statistic was used for all further tests. However, the results indicated that the coefficient estimates and standard error estimates were slightly biased and accordingly the type I error estimates of the Wald test were inflated/deflated when testing the effect of SNVs with very low MAF when there was a moderate or strong dependence between the traits (Kendall's $\tau \geq 0.5$), see Supplementary Tables A.13-A.14. Histograms of the Wald test statistics are shown in Supplementary Figure A.3 for Kendall's $\tau = 0.2, 0.5, 0.8$ and MAC $= 1, 2, 4, 10$. The underlying reason for this inflation was that the asymptotic distribution assumption for the Wald test did not hold when testing the association of SNVs with very few minor alleles (especially for MAC $= 1, 2$) under a finite-sample setting for Kendall's $\tau \geq 0.5$.

In order to obtain an approximation to the finite sample distribution of the test statistic, the Wald test statistics were adjusted using the empirical approach described in the following. Adjusted coefficient and standard error estimates could be obtained accordingly. First, $m = 1,000$ independent simulated sets of phenotypes were obtained under the null hypothesis as described in section 4.2.1, but conditioning on only one SNV without other covariates, for 251 different SNVs with MAC between 1 and 958 (which equal MAFs of $0.0005, 0.001, 0.0015, \ldots, 0.4790$ respectively). Next, C-JAMP point estimates $\widehat{\beta}_{XY}$ and their standard error estimates $\widehat{SE}(\widehat{\beta}_{XY})$ were obtained and the Wald test statistic was computed by $W = \frac{\widehat{\beta}_{XY}}{\widehat{SE}(\widehat{\beta}_{XY})}$. Standard normal Q-Q plots of these Wald test statistics suggested that the actual finite-sample distribution is a non-standard normal distribution which can be obtained as a linear transformation of the standard normal quantiles. Hence, in order to obtain adjusted Wald test statistics that follow a standard normal distribution, first a linear regression of 1,000 (sorted) standard normal quantiles Z on the (sorted) 1,000 empirical Wald test statistics W was computed for each Kendall's $\tau = 0.2, 0.5, 0.8$ and MAC $= 1, \ldots, 958$:

$$Z = \alpha_{\tau,\text{MAC}} + \beta_{\tau,\text{MAC}} \cdot W + \varepsilon. \tag{A.4.2}$$

These estimates can then be used to obtain the adjusted Wald test statistics W_{adj} based on

$$W_{adj} = \widehat{\alpha}_{\tau,\text{MAC}} + \widehat{\beta}_{\tau,\text{MAC}} \cdot W. \tag{A.4.3}$$

The estimates, $\widehat{\alpha}_{\tau,\text{MAC}}$ and $\widehat{\beta}_{\tau,\text{MAC}}$ from (A.4.2) are shown in Supplementary Figure A.4, and illustrate again that there aren't any problems for Kendall's $\tau = 0.2$ but that there is some bias for SNVs with small MAF and Kendall's $\tau \geq 0.5$. Next, these estimates were used to compute adjusted Wald test statistics for $\tau = 0.5, 0.8$ of the main simulation study results in Supplementary Figure A.3. Histograms of the adjusted Wald test statistics are shown in Supplementary Figure A.5 and indicate that all biases have been removed. Consequently, the adjusted Wald-test statistics were used for all further analyses in the simulation study. In

order to obtain an efficient generalization of this Wald test statistic adjustment which can be used in the analysis of any other dataset, the following approach can be used: Based on the estimated intercept and slope estimates of equation (A.4.2) in Supplementary Figure A.4, the functional form of the relationship between MAC and Kendall's τ with intercepts and slopes can be estimated. The following functions could be used to explain the relationship between MAC and intercept/slope:

$$\text{Intercept}_{\tau=0.5,0.8} = -\frac{1}{(A + B \cdot \text{MAC})^C}, \tag{A.4.4}$$

$$\text{Slope}_{\tau=0.5} = 1 + \frac{1}{(A + B \cdot \text{MAC})^C}, \tag{A.4.5}$$

$$\text{Slope}_{\tau=0.8} = 1 - \frac{1}{(A + B \cdot \text{MAC})^C}. \tag{A.4.6}$$

Fitting these functions to the data can be done using the least-squares estimation method and the `optimx()` function in the `optimx` R package. The parameter estimates of A, B and C are shown in Supplementary Table A.15, the estimated functions in Supplementary Figure A.6, and histograms of the adjusted Wald test statistics using this functional approach in Supplementary Figure A.7. They indicate a good fit to the points, yielded a similarly acceptable adjustment of the Wald test statistics, and could be used in any application.

A.4.2.2 Empirical type I error of C-JAMP

In a first additional investigation of type I errors, C-JAMP was investigated for testing the association of a gene with one trait Y_2, when the two traits Y_1 and Y_2 are correlated, but the genetic effects are only on the other trait Y_1. The results (based on adjusted Wald test statistics) showed that for a weak or moderate dependence between the traits, type I errors are valid (and rather deflated) for association tests with the non-associated second trait (Supplementary Table A.18). However, for the situation that there is a strong dependence between the traits but the gene is only associated with one trait, there was a slightly increased empirical type I error for testing the association with the second trait. Further investigations indicate that the reasons are slightly biased point estimates and slightly higher standard error estimates in this situation compared to the null model when both traits are unrelated with the gene (data not shown). Hence, in this situation, the Wald test statistic adjustment is not sufficient to estimate the correct finite sample reference distribution. This situation can be checked in application to real data, by conducting additional tests based on univariate models of traits as a sensitivity check.

The second additional investigation regarding type I errors concerns the situation when the dependence structure is misspecified in C-JAMP. The results of the main simulation study described here are based on using the Clayton copula for both generating the phenotypes as well as for the analysis. For practical applications, the 2-parameter copula (which contains the Clayton copula as a special case) encompasses a large variety of both symmetric and non-symmetric dependence structures so that it can be a good fit for many different practical datasets. Alternatively, other copula functions could be used, for example the Gaussian copula for fitting multivariate normal dependence structures. For an investigation of the empirical type I errors when the dependence structure is misspecified, the performance of C-JAMP based on the 2-parameter copula is evaluated when the data is generated from a

bivariate normal dependence structure. The same genetic data as in the main simulation study was used, as well as the same marginal models for the traits Y_1 and Y_2, but they were generated from a bivariate normal distribution, conditional on the covariates z_1, z_2:

$$\begin{pmatrix} Y_1|z_1,z_2 \\ Y_2|z_1,z_2 \end{pmatrix} \sim N(\mu,\Sigma), \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

where μ_1, μ_2 are $E(Y_1|z_1,z_2), E(Y_2|z_1,z_2)$ in the marginal models, respectively. Here, ρ was specified to generate a weak, moderate, or strong dependence between the traits (i.e., so that Kendall's $\tau = 0.2, 0.5, 0.8$) based on the relation $\rho = sin(\frac{\tau \cdot \pi}{2})$. The empirical type I error levels of C-JAMP are shown in Supplementary Table A.17. They show that the nominal levels are all close to the nominal levels for all considered situations based on testing the genetic effect with the unadjusted Wald test statistics. Hence, they would be slightly inflated for moderate dependence between the traits if the adjusted Wald test statistics were used (and slightly deflated for $\tau = 0.8$). It should be generally always checked in a first step, whether the dependence structure is appropriately captured by the used copula family.

A.4.3 Simulation study: CIEE - Discussion of the G-estimation approach for time-to-event traits

Lipman et al. (2011) proposed a method to detect direct genetic effects on a target time-to-event phenotype in the presence of genetic associations with an intermediate phenotype that affects the target time-to-event phenotype. The method is based on the methodology introduced in Vansteelandt et al. (2009) where the target phenotype was a completely observed continuous variable, which was modeled by a linear regression model. To estimate and test the direct effect of a genetic marker on the target phenotype, Vansteelandt et al. (2009) presented a two-stage method which includes an adjustment stage to remove the effect of an intermediate phenotype on the target phenotype. This method uses the principle of the sequential G-estimation (Goetgeluk et al., 2008) for making inference on direct effects in linear models. In the first stage, the effect of the intermediate phenotype on the target phenotype is estimated and an adjusted target phenotype is obtained by removing the effect of the intermediate phenotype. In the second stage, inference on the direct effect of the genetic marker on the target phenotype is obtained by regressing the adjusted target phenotype on the genetic marker.

The two-stage method works effectively under a linear regression model with a completely observed outcome. The goal of Lipman et al. (2011) was to extend the two-stage method to the setting in which the target phenotype is a time-to-event variable and subject to censoring. They considered PH regression and AFT models of a time-to-event phenotype, and proposed a new adjustment method to remove the effect of the intermediate phenotype on the target time-to-event phenotype. They obtained the adjusted survival times by subtracting the mean of observed survival times and the so-called "partial" Deviance residuals from the observed survival times (equation (4) in Lipman et al., 2011). In the following, we show that this adjustment method does not remove the effect of the intermediate phenotype on the target time-to-event phenotype.

We assessed whether the effect of the intermediate phenotype is removed from the defined adjusted target time-to-event phenotype given in equation (4) of Lipman et al. (2011). To

this end, we conducted a Monte Carlo simulation study. The design of our simulation study is similar to that of Lipman et al. (2011), but to be able to explain our points easily, we considered the simpler causal diagram in Supplementary Figure A.8. We generated 1,000 replicates, each of sample size 1,000. The genotype X was sampled from the Bernoulli distribution with mean 0.25, representing a dominant genetic effects model, and the intermediate phenotype K from the normal distribution with mean $0.5+1.0x$ and variance 1. Furthermore, we generated the target phenotype (i.e., survival times T) from the PH regression model with hazard function $h(t|k,x) = h_0(t)\exp(\beta_1 x + \beta_2 k)$, where $h_0(t)$ is the Weibull baseline hazard function with a unit scale parameter and shape parameter λ= 1.0 or 1.5. We considered a simple scenario, where survival times are all observed and there is no censoring. This simple setup allowed us a better examination of whether the adjustment method of Lipman & Lange (2011) works.

After applying the adjustment procedure in Lipman et al. (2011) to each replication sample, we investigated whether the intermediate phenotype K has any effect on the adjusted target phenotype $\tilde{T}$. To be able to assess the effect of K on $\tilde{T}$, one option is to fit a plausible parametric conditional model for $\tilde{T}$ given K and X (i.e., $\tilde{T}|K,X$). Based on the assumption that the effect of K on $\tilde{T}$ is removed, Lipman et al. (2011) suggested to use a standard linear regression model of $\tilde{T}|X$ in their equation (5) to estimate the direct effect of X on T. However, there is no discussion why this model was used. The appropriateness of a linear regression model of $\tilde{T}|K,X$ is in question. Nevertheless, we first considered the linear regression model of $\tilde{T}|K,X$ as it was suggested. Under this model assumption, if the adjustment method successfully removes the effect of the intermediate phenotype K on T, the coefficient of K will be zero in the linear regression model of $\tilde{T}|K,X$. Supplementary Table A.27 displays the empirical means and standard deviations of least square estimates of the coefficient of K over 1,000 replicates, which shows that the effect of K on the adjusted target phenotype $\tilde{T}$ has not been removed.

However, when we checked the linear regression assumptions, we observed that the linearity and homoscedasticity assumptions were violated. We also noticed that it is difficult to confirm a plausible distribution and a plausible model for $\tilde{T}|K,X$. Therefore, we assessed the association between the adjusted phenotype $\tilde{T}$ and the intermediate phenotype K using a nonparametric association measure instead of assuming a parametric form to model the causal relationship. We used a rank-based association measure, Kendall's τ, to assess the association between $\tilde{T}$ and K. Supplementary Table A.27 also shows the proportion of p-values that are below the nominal value of 0.05 for testing the null hypothesis that K is not associated with $\tilde{T}$ (i.e., $H_0 : \tau = 0$) for each stratum of X. We observed that there is still a strong association between the adjusted target phenotype $\tilde{T}$ and the intermediate phenotype K and the association becomes stronger as the effect of K on T, β_2, increases.

Contrary to what is proposed in Lipman et al. (2011), we showed here that the effect of the intermediate phenotype is not removed from the adjusted target time-to-event phenotype. Hence, the direct genetic effect on the target time-to-event phenotype cannot be estimated using the proposed approach under the causal directed acyclic graph. In addition, there is no discussion on whether $\tilde{T}$ is adjusted for censoring and how a standard linear regression model without considering censoring (as in equation (5) in Lipman et al., 2011) could be used to

estimate the direct effect of a genetic marker X on the time-to-event variable T. Even when there is no censoring, modeling the adjusted target time-to-event phenotype $\tilde{T}$ by using a standard linear regression model of $\tilde{T}|X$ as in equation (5) in Lipman et al. (2011) is not a valid approach, as is using the test statistic for testing the absence of a direct genetic effect on the target time-to-event phenotype.

A.4.4 Detailed materials and methods description of the obesity study

A.4.4.1 Assessment of fat mass and fat distribution

As classic anthropometric measures, waist circumference (WC), height, and weight were taken by trained personnel. BMI was calculated as weight (in kg) divided by squared height (in m), and waist-hip-ratio (WHR) as the ratio of WC and hip circumference. Whole-body MRI scans excluding head and arms were performed in Potsdam on Siemens 1.5T Avanto scanners using a 2-point Dixon technique with a 3D gradient echo sequence, and body fat measures were obtained using a fully-automated segmentation approach including image-processing methods and statistical shape models (Wald et al., 2012). The technical details of the approach as well as the protocol, sample ascertainment, repeatability and reproducibility of the measurements, and association with standard anthropometric measures have been described in detail elsewhere (Wald et al., 2012, Neamat-Allah et al., 2014). Available measures include the amount of VAT (in the abdominal cavity, i.e., around and between the organs in the abdomen), SAT (fat tissue under the skin), and CAT (fat tissue around the heart/ heart vessels) compartments, with SAT also being measured in the abdomen and thigh. In addition, total mass of skeletal muscle, bone marrow and total body volume (TBV) were assessed, and total AT (TAT) was calculated as the sum of AT in all body compartments. Furthermore, the ratios $\frac{\text{TAT}}{\text{TBV}}$, $\frac{\text{VAT}}{\text{TBV}}$, $\frac{\text{SAT}}{\text{TBV}}$, $\frac{\text{VAT}}{\text{TAT}}$, $\frac{\text{SAT}}{\text{TAT}}$, $\frac{\text{SAT}}{\text{VAT}}$, were obtained as relative measures of AT mass and of AT distribution. Previous studies showed that AT quantification was highly reproducible with a coefficient of variation of about 0.3% for SAT and TAT, and 3.5% of VAT (Wald et al., 2012).

The variables including the quantification of absolute and relative fat mass were already cleaned and quality-controlled in previous studies, and there was no indication for any further problems or errors. SAT had a slighty skewed distribution and was log-transformed. See Figures A.9-A.10 for a visualization of the distribution of SAT and $\frac{\text{SAT}}{\text{TAT}}$ (stratified by gender).

A.4.4.2 Subcutaneous fat tissue biopsies

Of the 673 eligible participants for AT biopsies, 186 refused participation and 209 were excluded (mostly due to reported use of anticoagulant or antiplatelet medication within the last seven days). Further exclusion criteria were any form of prevalent coagulopathy, an allergy against lidocain or plaster or a disposition for keloids were excluded from the fat biopsy. The goal was to extract 0.4g or more of abdominal SAT with a needle aspiration method. The biopsy was conducted by a specially trained physician, assisted by a nurse. Subcutaneous anesthesia was achieved by infiltrating a skin area of approximately 5x5cm with 10mL of 1% lidocaine. Afterwards, a 14G Strauss needle was inserted into the SAT and using a syringe, approx. 4mL of sterile physiological saline solution was injected into the SAT.

Afterwards, the plunger was fully retracted to create negative pressure in the syringe and to aspirate tissue. With the plunger in this position, the needle was moved in all directions under the skin two to four times. The syringe was then removed with the plunger pulled back and the aspirate was transferred to a 50mL tube containing 12.5mL of 10% sodium citrate solution. This procedure was repeated ten times. The tube with the fat, blood, saline and sodium citrate solution conglomerate was then transferred to the laboratory and further processed within 30 minutes. Larger blood components were removed manually using a 10mL syringe. Then, sterile physiological saline solution was added and the tube centrifuged at 100g for 2 minutes. The fluid below the AT was then removed with a syringe. This procedure was repeated until no more blood components were visible. The AT was then transferred to a cryotube and centrifuged again for 2 minutes at 100g. After removing the fluid below the AT, saline solution was added again and the tube centrifuged for ten minutes at 100g. Residual fluids were extracted using a syringe and the remaining AT was weighted and then frozen at -80°C. 273 of the performed 278 biopsies yielded material, with on average 0.88g aspired AT (mean=0.05g, max=3.51g).

A.4.4.3 RNA extraction from SAT Biopsies

From the SAT biopsies, the total RNA, genomic DNA, and total protein from the fat tissue samples were purified and separated using the Qiagen All Prep DNA/RNA Mini Kit. The purified genomic DNA has an average length of 15-30 kb. Regarding the RNA, only RNA molecules longer that 200 nucleotides were purified and with the employed standard protocol, all short RNA molecules with length less than 200 nucleotides were removed. These removed small molecules include most of the ncRNA and short mRNA. The quantity and integrity of the purified DNA and RNA was verified using the NanoDrop 1000 Spectrophotometer V3.7 (PeqLab) and the 2100 Bioanalyzer (Agilent), which uses an on-chip electrophoresis. For the assessment of gene expression of candidate adipokines with PCR, 2μg RNA were reverse transcribed to cDNA using the High Capacity cDNA Reverse Transcription Kit (Applied Biosystems).

A.4.4.4 Library preparation & multiplexing

Next, the extracted RNA was prepared for sequencing using the TruSeq RNA and DNA Sample Preparation Kit v2 (Illumina). First, it was polyA-selected to purify and enrich for mRNA, fragmented into small pieces and primed (with random hexamers) for cDNA synthesis. The cDNA products were then enriched with PCR and ligated to adapters in order to index different samples ("multiplexing"). In the ligation, single indexes were used. Finally, the cDNA libraries were created, validated, normalized (so that they had equal volume) and pooled. The resulting pooled single-indexed paired-end libraries of different samples were then applied to the flow cells (containing 8 lanes) on cBot (Illumina), so that multiple samples could be sequenced together.

A.4.4.5 Sequencing and demultiplexing

Next, the multiplexed probes were sequenced on the Illumina HiSeq 2000 platform in 201 sequencing cycles. For some intuition, the generated clusters are sequenced simultaneously, and the nucleotide in the sequence is obtained by capturing its fluorescence signal with a built-in camera. 198 probes were sequenced with the depth 6 samples per lane, and 2 probes were sequenced each on one lane as a sensitivity check and to assess how many more genes can be (reliably) detected with a greater sequencing depth. Of the 198 probes, there were technical problems on one lane on one flow cell so that about 20% of the data from the first base call of these 6 probes was missing. This run was repeated and the partly missing sequences can be used for reliability checks. The raw data contains the sequences of quality-scored base calls and was saved in .bcl files. Next, the `CASAVA` (Illumina) software was used for converting the .bcl files to .fastq files. In the same step, the multiplexed samples were demultiplexed and the raw reads of each sample are extracted and saved.

A.4.4.6 Quality control of the demultiplexed raw sequencing reads

At first, the quality metrics from the demultiplexed fastq files were inspected using the `bcl2fastq` software (Illumina). Supplementary Figure A.11 gives an overview about the number of raw reads of all probes, where the reads of the 6 re-sequenced probes are aggregated. The mean number of raw reads (of all probes except the 2 deeply sequenced probes) was 64,095,856 (SD=7,518,970), with minimum 43,373,110 and maximum 85,591,020 raw reads.

For an assessment of the sequencing quality, Illumina uses the Q-score as quality scoring method, which is a prediction of the probability of an incorrect base call. More specifically, given a base call X, the quality score $Q(X)$ is computed as $Q(X) = -10\log_{10}(P(\neg X))$ where $P(\neg X)$ is the estimated probability of the base call being wrong. Hence, a quality score of 10 indicates an error probability of 0.1, a quality score of 20 indicates an error probability of 0.01, a quality score of 30 indicates an error probability of 0.001 etc. All probes had a high percentage (on average 88.0%, SD=1.9%) of bases with high quality ($Q > 30$), see Supplementary Figure A.12. Only the first 6 sequenced probes had a slightly lower percentage around 80%, and all other probes had a percentage between 85% and 93%. The mean quality score of bases in a probe was 32.5 (SD=0.5) and showed a very similar pattern (Supplementary Figure A.13).

Next, more details of the quality checks are reported by investigating the sequence quality, GC content, sequence base content, the presence of adapters and duplicated reads using the `FASTQC` tool (Andrews, 2010). The sequence quality was high for all probes, across all base pair positions of all reads, see Supplementary Figures A.14-A.15 with one representative probe as example. The GC distribution across all reads was close to the expected theoretical distribution for all probes, with only minor deviations from some probes which didn't indicate systematic biases or deviations, see Supplementary Figure A.16 for a representative illustration of one probe. The sequence base content was not random for the first 10 bases across all reads (see Supplementary Figure A.17), but this was observed for all probes as a usual artefact due to the (non-)random hexamer priming, and is not considered a reason for concern. Next, there was no indication for any problems with adapters (Supplementary

Figure A.18). Due to the PCR-steps involved in the cluster generation in RNA-Seq, duplicate sequences were naturally observed (Supplementary Figure A.19), around 50% for all probes. There were not any specific sequences that were reported to be systematically duplicated, hence there don't seem to be any systemic problems.

As a summary, the quality control checks of the raw reads showed a consistently high read quality without indication for systematic sequence biases, presence of adapter sequences, or duplication levels beyond what can be expected from RNA-Seq. Hence, the fastq files were parsed to the alignment stage with `TopHat2` without trimming or deleted of reads.

A.4.4.7 Read alignment

The reads were aligned to the human reference genome GRCh38 (Homo_sapiens.GRCh38.78) using `TopHat2` (version 2.0.12) with `Bowtie2` (version 2.0.6.0) and `samtools` (version 0.1.18.0), which maps the RNA reads in the presence of insertions, deletions and gene fusions. Since `Bowtie2` has limitations with respect to alignments containing large gaps, `TopHat2` addresses these issues so that, for example, noncoding RNAs can be mapped as well as transcripts spanning introns. `TopHat2` uses the .bcl files and generates .bam files with the aligned reads. In the alignment, all reads were used without trimming or discarding reads with low-quality calls.

A.4.4.8 Quality control of the aligned reads

In the following, descriptive statistics of the alignment quality metrics are investigated. Supplementary Figures A.20-A.21 show the number of mapped reads in absolute and relative frequencies, and Supplementary Figures A.22-A.23 the number of aligned pairs in absolute and relative frequencies. In these plots, the reads of the 6 re-sequenced probes are aggregated. On average, 90.3% (SD=1.2%) of all reads could be aligned. Furthermore, on average, 85.7% (SD=1.9%) or all raw reads could be aligned in pairs and on average 91.8% (SD=1.3%) of the mapped reads had a high mapping quality (Supplementary Figures A.24-A.25). These plots and metrics illustrate that in addition to the high sequencing quality, the alignments had high quality.

For additional investigations, alignments using `TopHat` (Trapnell et al., 2009) or `TopHat2`, `Bowtie` (Langmead et al., 2009) or `Bowtie2`, and alignments to GRCh37 were generated and inspected. In comparisons, it was observed that (i) alignments using `TopHat2` with `Bowtie2` provided more aligned reads and more aligned reads with high mapping quality compared to `TopHat` with `Bowtie2`, which in turn provided more (high-quality) aligned reads compared to `TopHat2` with `Bowtie`, (ii) alignments to GRCh38 yielded more aligned reads (but also more multiple alignments) compared to alignments with GRCh37.

A.4.4.9 Read counts as raw measures of gene expression levels

The aligned reads were first sorted using `samtools` and then `htseq-count` was used to obtain counts as a raw measure of the gene expression levels. Counts were obtained for genes with respect to ensembl gene identifiers. Default settings were used to discard aligned reads with mapping quality smaller than 10. Since all mapped reads had either mapping quality higher

than 30 or lower than 10, only high-quality-mapped reads were used to obtain counts. Reads overlapping multiple genes were not counted for any gene (based on the default mode "union").

For an illustration of the amount of filtered probes (reads which were not properly paired, of low mapping quality, or where the reads or mates had multiple alignments), Supplementary Figure A.26 shows the absolute number of mapped and counted reads (median 23,850,000 counted reads per probe), and Supplementary Figure A.27 shows the percentage of mapped and counted reads relative to the raw reads, which was on average 56.7% (SD=3.0%).

A.4.4.10 Quality control and normalization of read counts

The ensembl database (Aken et al., 2016; Hubbard et al., 2002) lists in total 64,253 genes for the reference genome GRCh38.78 (obtained from `http://dec2014.archive.ensembl.org/index.html`). Based on the obtained raw gene expression levels, 48,126 genes were expressed in at least one probe. Interestingly, only 107 of the 48,126 genes ($\approx 0.2\%$) were uniquely observed in the 2 deeply sequenced probes which yielded 6 times as many reads. On average, the expression of 27,690 genes was observed per sample (SD=1,274, min=25,430, max=33,280), 19,460 genes had at least 5 counts per sample (SD=1,042, min=17,600, max=25,910), 17,500 genes had at least 10 counts per sample (SD=940, min=15,890, max=23,580), and 13,630 genes had at least 50 counts per sample (SD=696, min=12,420, max=18,470). An illustration of the distribution of these numbers is shown in Supplementary Figure A.28. From another perspective, Supplementary Figure A.29 shows histograms of the average number of observed counts per gene, contrasting the changes when probes were sequenced at the depth of 1 sample per lane, 3 samples per lane (aggregating the 6 re-sequenced probes), and 6 samples per lane. As can be expected, the number of genes without observed counts decreases, and the average count per gene as well the maximum number of counts per gene increases, but the shape of the average observed counts per gene and the overall pattern don't change with sequencing depth.

For further details regarding the comparison of different sequencing depths, Supplementary Figure A.30 shows the relative frequency of gene classes that are detected with the different sequencing depths, for one probe sequenced at 1 sample per lane and one probe sequenced at 6 samples per lane. In the left panel, the percentage of detected features (e.g., protein coding genes) in each probe is shown. As could be expected, this percentage is always higher for the deeper sequenced probe, but again, the differences are not substantial especially for protein coding genes. The right panel shows the relative abundance of genes within each sample and, for example, that between 50% and 60% of detected genes are protein coding. Supplementary Figure A.31 shows the same figures contrasting the two deeper sequenced probes, and doesn't indicate any differences.

In the analysis of gene expression measures from RNA-Seq, the observed counts are often dependent on the gene length (the longer the gene, the higher the counts), which can affect downstream analyses (Hansen et al., 2012). Hence, instead of the raw counts, normalized gene expression measures should be used. Popular measures include CPM (counts per million: counts scaled by the number of genes per million), RPKM (reads per kilobase of exon model per million mapped reads: counts per total counts in million and per gene length in kilobases; Mortazavi et al., 2008), and TPM (transcripts per million: counts per gene length adjusted

for total counts per gene lengths in million; Li & Dewey, 2011). All are within-sample normalizations, and RPKM as well as TPM correct for gene length and library size (i.e., total number of reads). Advantages of TPM are that it allows for between-sample comparisons (Wagner et al., 2012) and it was chosen as normalization measure here.

TPM-normalized counts were computed for the 48,019 genes which had non-zero counts in at least 1 sample as follows: Let d denote the number of genes under investigation, Count_s their observed counts, and $\text{Length}_s, s = 1, \ldots, d$ their gene length. Then, for a given gene s:

$$\text{TPM}_s = \text{RPK}_s \cdot \frac{10^6}{\sum_{s=1}^{d} \text{RPK}_s},$$

where $\text{RPK}_s = \text{Count}_s \cdot \frac{10^3}{\text{Length}_s}$ are the read counts divided by the gene lengths per kilobase. The gene lengths were obtained from the ensembl database through `biomaRt` (Durinck et al., 2009) at `dec2014.archive.ensembl.org` (accessing GRCh38.78).

In addition to the within-sample normalization, the TMM (Robinson & Oshlack, 2010) method was used for a between sample normalization to account for potential sequencing biases. TMM computes the trimmed mean of M-values between each pair of samples and thereby normalizes for RNA composition using scaling factors for the total number of reads (i.e., library sizes), which minimizes the log-fold changes between samples. The effective library size is computed and used instead of the observed library size.

It has been shown that the GC-content of a gene mentioned in section A.4.4.6 can bias downstream analyses (Hansen et al., 2012; Risso et al., 2011). Hence, it was investigated for the TMM-normalized TPM values (see Supplementary Figure A.32). The shown scatterplots of the mean gene expression (i.e., TMM-normalized TPM value over all samples) versus the GC content of all genes don't indicate an association, and a linear regression confirmed that there wasn't any (linear) association. As a result, the gene expression measures were not further normalized for GC content in order not to remove true biological variance and the TMM-normalized TPM values were used as primary measure of gene expression. All genes with observed counts for at least one individual were passed on to the following stages of the data processing.

For an illustration, the distribution of the TMM-normalized TPM values is plotted in Supplementary Figure A.33 for the candidate genes investigated in the study described in section A.4.5. This shows that the distribution is normally distributed for some genes, but very skewed for others. Supplementary Figures A.34-A.35 illustrate the differences between TMM-normalized and non-normalized TPM values, which shows that they are very similar and confirms that there are few biases between experiments (i.e., between samples).

As further validity checks, the correlation between gene expression measures obtained from RNA-Seq and PCR for 6 candidate genes has been described in section 5.2.2.1. See Supplementary Figure A.36 for an illustration of these associations. In addition to these results, the correlations between the TMM-normalized TPM values and PCR gene expression were higher or equal compared to using other measures of RNA-Seq gene expression (e.g., TMM-normalized or raw RPKM or CPM values, or raw counts), supporting the use of TMM-normalized TPM values as gene expression measure for the main analysis.

A.4.4.11 Transformations of normalized read counts

Next, the distribution of the gene expression measures was investigated. First, the skewness distribution of all genes with non-zero counts in at least 1 of the 198 regularly sequenced probands is shown in Supplementary Figure A.37. It shows that there are few genes whose gene expression had a left-skewed distribution (minimum skewness = -1.1). The gene expression of many genes was slightly skewed, but about half of all genes had a strongly right-skewed distribution. Transformations (such as a log-transformation) lead to problems for very lowly expressed genes, where almost everyone has an observed count of 0. Hence, the following analyses and transformations were restricted to those 30,917 genes, where at least 25% of the people (i.e., at least 50 probes) had non-zero observed counts. The skewness distribution of these genes is shown in Supplementary Figure A.38, and confirms that most lowly expressed genes had a very skewed distribution. Of the 30,917 genes, 15,569 had skewness greater than 1, and 5498 had skewness greater than 2.

Next, a log-transformation was applied to those genes whose gene expression had a distribution with skewness greater than 1. For these genes, before the log-transformation, the value 0.000001 was added to the (TMM-normalized TPM) gene expression to ensure that only positive values were log-transformed (the minimum non-zero gene expression value in all genes was 0.00014). The skewness after this transformation is shown in Supplementary Figure A.39, and shows that while there were no right-skewed genes anymore, the result is not satisfactory since now there were consequently many left-skewed genes.

As a result, another approach was pursued, and the skewed genes (i.e., with skewness greater than 1) were transformed using the Yeo-Johnson transformation (which is the Box-Cox transformation of U + 1 for nonnegative values, and of $|U|+1$ with parameter $2-\lambda$ for U negative) using the `yjPower()` function in the `car` R package (Fox & Weisberg, 2011), where the parameter λ for the transformation is determined in a first step using the `powerTransform()` function. This approach seemed to be successful in removing the skewness for all genes, as shown by the histogram of the skewness of all genes in Supplementary Figure A.40, and histograms of the gene expression of the first 50 genes in Supplementary Figures A.41-A.42. Still, 7,953 genes didn't seem to be normally distributed according to the Kolmogorov-Smirnoff test with a p-value smaller than 0.001, for example, due to bimodal or other forms of distributions. However, for those genes, there wasn't a clear indication which transformation could yield normally distributed measures. Hence, the Yeo-Johnson transformed gene expression measures were used as final measure for the analysis. To incorporate that the normality of some genes is questionable, robust analyses incorporating multiple genes (e.g. the GO term enrichment analysis) were performed, and top associated genes with highest significance in the main analyses were inspected manually.

A.4.4.12 SNV calling

In order to investigate genetic variants, SNVs (in coding regions) were called from the RNA-Seq .bam files using `samtools` and only correctly paired reads. In more detail, `samtools mpileup` and `bcftools view` were used to call the SNVs and generate the base call format (.bcf) and variant call format (.vcf) files. All samples were used as input simultaneously call the genotypes of all observed SNvs for all samples, by using priors and the information

from the other samples for samples where the SNV was not observed. This also yields a higher calling quality. For a better computational efficiency, the genome was divided into 550 regions (including the X and Y chromosomes, MT DNA, and scaffolds of unknown position), and SNVs were called separately for the regions. In the calling, no filters regarding depth or calling quality were used, and these quality control steps were done in the following. In total, 2,029,767 biallelic SNVs could be called as well as 79,618 multi-allelic (>2 allelic) SNVs and INDELS. The 2,029,767 biallelic SNVs included 1,978,091 autosomal SNVs, 45,874 SNVs on the X and Y chromosomes, 760 mitochondrial SNVs, and 5,042 referenced SNVs on scaffolds.

All analyses in the following focus on the autosomal SNVs. The separate vcf files were aggregated by chromosome, and the next filtering steps (by depth and quality) were done on the chromosome level (also for higher computational efficiency). More specifically, SNVs were filtered if the read depth was smaller than 100 (based on high-quality reads across all 200 samples), if their depth was larger than 3 times the average depth (possible PCR amplification bias) and if their quality score was smaller than 15. After these basic filtering steps, 538,506 (autosomal) SNVs were retained. Deletion of monomorphic SNVs yielded 531,397 SNVs, and after deleting SNVs which had a p-value smaller than 10^{-7} in the exact test of the Hardy-Weinberg equilibrium (computed using the `genetics` R package; Warnes et al., 2013), 509,009 SNVs were available for the main analysis. Across all SNVs and samples, the average coverage per position per sample was 6.6. Next, genotypes (0, 1, 2) were computed, MAFs were calculated, and the genotypes of those SNVs with MAF>0.5 were reverted and the MAF recomputed. For an illustration of the MAF distribution of all 509,009 SNVs, see Supplementary Figure A.43.

A.4.4.13 Assessment of gene expression of candidate adipokines with PCR

After extracting the RNA and converting it to cDNA, quantitative real time PCR was used to evaluate the gene expression in SAT of 9 target genes: adiponectin (AdipoQ), leptin (LEP), soluble leptin receptor (sOB-R), resistin (RETN), fatty acid binding protein 4 (FABP4), tumor necrosis factor alpha (TNFα), interleukin 6 (IL6), as well as 11β-hydroxysteroid dehydrogenase type 1 and type 2 (11β-HSD1, 11β-HSD2). qPCR was performed using the Applied Biosystems 7500 Fast Real-time PCR system with TaqMan technology (ABI, Darmstadt, Germany). The two-step PCR conditions were 20s at 95°C, 40 cycles with 3s at 95°C, and 30s at 60°C (5μL reaction volume, 4ng template). In each amplification cycle, a cycle threshold (Ct) value was obtained. For each probe and for each gene, gene expression was measured in triplicates and the three Ct values were averaged for each individual to gain a more robust measure. In addition, the expression of the reference rRNA housekeeper gene 18S was measured for each probe on the same plate for four genes (adiponectin, resistin, IL6, TNF-α), also in triplicate, and the three Ct values were averaged as well. These four 18S average Ct values were very similar across plates, and further averaged for each probe to obtain the reference 18S expression (in form of a grand average Ct value) for each probe. Relative normalized gene expression levels ΔCt in comparison to the reference gene were obtained for each gene and each probe, from the difference between the averaged candidate gene Ct value and the 18S grand averaged Ct value. As measure for gene expression in the analysis, $2^{-\Delta Ct}$ values were used, assuming that the number of amplified target molecules at the threshold cy-

cle is identical for the candidate genes and the housekeeper (Livak & Schmittgen, 2001). Gene expression of one sample was significantly different from all other probes for all genes and the housekeeper, so that the probe was discarded from further description and analysis, yielding a sample size of 199 for the analysis of gene expression. For resistin, SAT gene expression levels were very low leading to unobservable Ct values across triplicates for 2 samples, which had to be excluded from the analysis of resistin gene expression. The inter-assay coefficient of variation (across the four 18S Ct measures of each probe on the 4 different plates) was on median 1.2%. For all genes, the median intra-assay coefficient of variation (across the triplicate Ct measures of all sample probes on all 84 plates) was smaller than 1.5%.

A.4.4.14 Assessment of plasma concentrations of candidate adipokines

In the study, the plasma levels of the adipokines adiponectin, leptin, soluble leptin receptor (sOB-R), resistin, interleukin 6 (IL6) and fatty acid binding protein 4 (FABP4) were examined. Blood samples were collected during the reinvitation of participants from 2010-12, processed, divided into 1ml nunc tubes and stored in -80° C freezers. EDTA plasma probes were thawed, split into predetermined volumina necessary for the different ELISA kits and then further stored at -80°C. Plasma levels of total, high molecular weight (HMW), and HMW + medium molecular weight (MMW) adiponectin were measured using ELISAs from Alpco (Salem, USA), and concentrations of MMW and low molecular weight (LMW) adiponectin computed by subtraction. Leptin, sOB-R, resistin and FABP4 were measured using ELISAs from BioVendor (Brno, Czech Republic), and IL6 was measured using an ELISA from R&D Systems (Minneapolis, USA). All samples were measured in duplicates according to standard protocol on the TECAN Infinite 200 PRO reader (Männedorf, Switzerland). Concentration estimates adjusted for reference absorption and background noise were obtained using the observed absorption measures and a 4-parameter logistic (4PL) model (Ritz & Streibig, 2005), fitted with the `optimx` package in the statistical software R version 3.3.1.

Among the 3,136 measures of all biomarkers of all participants, 9 single absorption measures were missing and 16 measures yielded low absorption measures so that their concentration couldn't be estimated using the 4PL model. This included 6 measures for leptin, 1 for sOB-R, 5 for FABP4, 3 for total adiponectin, 4 for HMW adiponectin and 6 for HMW+MMW adiponectin. For these low absorptions, concentrations were set to half of the respective minimally possible detected concentration so that after averaging duplicate measures, concentrations were missing for 1 sample for leptin, LMW and MMW adiponectin. The median intra-assay coefficient of variation (across the duplicate measures of all sample probes on all 48 plates) was <5% for all biomarkers, except for sOB-R, where it was 11%. The inter-assay coefficient of variation (of the averaged estimated concentrations of the two measures of a pooled probe across all 48 plates) was 6% for leptin, 14% for sOB-R, 11% for resistin, 6% for IL6, 8% for FABP4, 4% for total adiponectin, 8% for HMW+MMW adiponectin, and 11% for HMW adiponectin.

A.4.5 Study to predict adipokine plasma levels based on body fat compartments and adipose tissue gene expression

A.4.5.1 Overview

Over the past years, a number of circulating adipokines have been identified and proposed as mediators for the association of obesity with chronic disease risk. However, it is unclear to what extent the amount of subcutaneous (SAT) and visceral (VAT) adipose tissue (AT) and gene expression levels in AT predict adipokine concentrations on a population level. In a cross-sectional analysis of 200 participants from the population-based EPIC Potsdam cohort study, measures of AT mass were obtained using anthropometry and magnetic resonance imaging. In addition, gene expression in SAT, collected by biopsy, and plasma concentration of the cytokines adiponectin, leptin, soluble leptin receptor, resistin, interleukin 6 and FABP4 were measured. The primary aim of this study was to investigate to what extent different imaging-based AT measures and gene expression predict the circulating levels of these adipokines. In addition, it was evaluated which AT measures are best representing the metabolic activity of adipose tissue. This measure was then further used in section 5.2.

For leptin and FABP4, 77% and 37% variance of circulating levels were explained by SAT as well as VAT mass and gene expression. Leptin showed stronger correlations with SAT compared to VAT mass, while there was little difference between AT compartments regarding their correlation with FABP4. For both adipokines, plasma concentrations were more strongly correlated with AT mass compared to gene expression. For the remaining adipokines, less than 16% of the variance was explained by SAT, VAT, and gene expression. The data suggest that except for leptin and FABP4, SAT, VAT and gene expression in AT predict adipokine concentrations only to a small extent. These data point to other predictors. Further, while our findings do not contradict a potential role in disease development, these adipokines are unlikely to account for a large proportion of the association between adiposity and disease risk observed in epidemiological studies.

A.4.5.2 Introduction

Obesity is an established risk factor for a number of chronic diseases. Investigations into the underlying molecular processes have identified circulating adipose tissue-derived biomarkers, so-called adipokines (Fantuzzi & Mazzone, 2007; Pischon, 2009), that are - among various pathways - involved in glucose and lipid metabolism, inflammation, insulin signaling, and energy homeostasis (Ouchi et al., 2011; Fasshauer & Blüher, 2015). A number of these adipokines have also been associated with risk of cardiovascular disease, diabetes, hypertension, and cancer, and have been proposed as mediators for the association of obesity with chronic disease risk (Pai et al., 2004; Norata et al., 2007; Terra et al., 2011; Aleksandrova et al., 2012a, 2012b). Surprisingly, the variance of plasma adipokine levels explained by the body mass index (BMI) or by waist circumference - which are currently used to define obesity or abdominal obesity, respectively - is relatively modest, around 36% for leptin and only up to 10% for most of the other adipokines (Pai et al., 2004; Pischon, 2009; Aleksandrova et al., 2012a). However, BMI and waist circumference are only crude measures of the amount of subcutaneous and visceral adipose tissue (SAT and VAT), which are the major compartments

of human adipose tissue (Ibrahim, 2010). Also, the degree of gene expression on the tissue level has often not been taken into account as further predictor of plasma levels (Fredriksson et al., 2006; Campbell et al., 2013).

Magnetic resonance imaging (MRI) allows direct quantification of SAT, VAT and TAT (Ross, 2003; Karlsson et al., 2013; Lee & Kuk, 2013), but there is inconsistent empirical evidence as to what extent imaging-based assessment of VAT, SAT or TAT predicts adipokine levels (Fox et al., 2007; Maahs et al., 2007; Won et al., 2009; Arnardottir et al., 2013). Further, only few studies have reported on the association of imaging-based adipose tissue (AT) mass with gene expression and circulating plasma or serum concentrations, have investigated different research questions, and have mostly focused on specific patient groups and on the analysis of one or two candidate adipokines (Goldstone et al., 2002; Sutinen et al., 2003; Berndt et al., 2005; Pietiläinen et al., 2006; Kursawe et al., 2010).

The aim of our study was therefore to examine to which extent circulating plasma adipokines levels can be predicted the amount of SAT, VAT (assessed by MRI), and by AT gene expression (measured in SAT). Our hypothesis was that direct quantification of SAT and VAT as well as AT gene expression explain a substantial amount of the variance in adipokine levels. As secondary aim, it was investigated which anthropometric measures show the strongest association with adipokine AT gene expression and plasma levels.

A.4.5.3 Methods and materials

In the analysis, the study sample described in section 5.2.1.1, MRI-based AT measures described in section A.4.4.1, gene expression levels in SAT of the candidate adipokines measured by PCR (see section A.4.4.13) and plasma concentrations of the candidate adipokine measured by ELISA (see section A.4.4.14) were used. Statistical analyses were performed using R version 3.3.1. All gene expression, plasma concentration, and MRI-based AT mass measures were log-transformed to yield normally-distributed measures for the analysis. For descriptive statistics, mean/standard deviation (SD), median/median absolute deviation (MAD), or frequencies of selected study characteristics are described. Partial Pearson correlation coefficients of the AT measures with gene expression and protein levels, adjusted for sex, age, physical activity and occupational training were calculated. In sensitivity analyses, we also calculated Spearman correlation coefficients and found very similar results (data not shown). Linear regression models were fitted for the primary question how much of the variance of plasma concentrations can be explained by gene expression, AT mass, SAT gene expression x SAT mass, and personal/environmental factors, measured by the adjusted R^2 accounting for the number of predictors in the model. Here, the interaction of SAT gene expression and SAT mass also carries the biological interpretation of the cell-based measure of transcriptomic activity multiplied by the amount of tissue (i.e., number and size of cells). In each part, a complete-data analysis of the respective maximally available sample size was performed and missing values were removed. This yielded sample sizes between 200 (e.g., of descriptive statistics for anthropometric traits and personal characteristics) and 149 (in the full regression model 14 in Supplementary Table A.39 including all plasma concentrations, gene expression levels, AT measures, and personal characteristics) in the reported analyses below.

A.4.5.4 Results: sample characteristics

Descriptive statistics of the study participants regarding their personal characteristics, disease prevalence, AT measures, and gene expression and plasma concentration of biomarkers are shown in Table 5.6 and Supplementary Table A.34. The sample contained slightly more women than men (57% women, 43% men), with an average age of 65.1years (SD=9.0 years), a mean BMI of 27.9 (SD=4.2), and a low prevalence of cardiovascular and cardiometabolic diseases. SAT was on average 20.1kg (SD=5.2kg) for women, and 14.8kg (SD=4.3kg) for men. Both plasma levels and gene expression showed variation between persons and sexes. For example, leptin plasma levels were on median 44.6μg/ml (MAD=25.2μg/ml) in women and 44.6μg/ml (MAD=25.2μg/ml) in men, and the MAD of gene expression was also about half of its median.

A.4.5.5 Results: Correlation of adipokine concentations with body fat compartments

On the level of plasma concentrations (Supplementary Table A.35), for leptin and FABP4, the highest correlations (adjusted for sex, age, physical activity, occupational training) were observed with TAT ($r = 0.80$, 95%CI = $[0.74; 0.85]$ and $r = 0.53$, 95%CI = $[0.41; 0.63]$), which were of similar size compared to SAT ($r = 0.79$, 95%CI = $[0.72; 0.84]$ and $r = 0.50$, 95%CI = $[0.37; 0.61]$), and larger compared to the correlation with VAT ($r = 0.59$, 95%CI = $[0.47; 0.68]$ and $r = 0.44$, 95%CI = $[0.31; 0.56]$). For sOB-R, IL6, total, HMW, and LMW adiponectin, and resistin, the correlation of each adipokine with SAT, VAT, and TAT was of similar size (around $r = -0.40$, $r = 0.31$, $r = -0.25$, $r = -0.25$, $r = -0.10$ and $r = 0.05$, respectively). For MMW adiponectin, the correlations with VAT were higher compared to SAT and TAT ($r = -0.32$ for VAT versus $r = -0.14$ for TAT and $r = -0.08$ for SAT). For leptin, MMW adiponectin, and FABP4, MRI fat mass measures showed stronger correlations with plasma levels compared to the traditional measures (BMI, WC, HC, and WHR), whereas for all remaining adipokines, correlations with MRI measures were in a similar range as their correlations with traditional measures. The correlations of plasma adipokine levels with BMI were generally in a similar range as their correlations with WC. CAT showed smaller or equal correlations compared to VAT, SAT, TAT for all investigated adipokines. Finally, the correlations of absolute fat measures (SAT, TAT, VAT) were very similar compared to correlations using relative fat measures (i.e., $\frac{\text{SAT}}{\text{TBV}}$, $\frac{\text{VAT}}{\text{TBV}}$, $\frac{\text{TAT}}{\text{TBV}}$) for most biomarkers (data not shown).

A.4.5.6 Results: Correlation of subcutaneous adipose tissue adipokine expression with body fat compartments

In general, correlations between adipokine gene expressions and AT measures (Supplementary Table A.36) were weaker as compared to the correlations between plasma concentrations and AT measures (Supplementary Table A.35). Leptin expression showed the strongest correlation with SAT ($r = 0.37$); whereas sOB-R, resistin, FABP4, adiponectin, and IL6 showed similar associations with SAT, VAT and TAT ($r = 0.06$ to 0.17, $r = 0.15$ to 0.20, $r = -0.27$ to -0.24, $r = -0.30$ to -0.25, and $r = 0.19$ to 0.26, respectively, see Supplementary Table A.36).

A.4.5.7 Results: Correlation of SAT gene expression with adipokine concentrations

SAT leptin gene expression was moderately correlated with leptin concentrations ($r = 0.52$, Supplementary Table A.37). Correlations between SAT gene expression and plasma levels were modest for total, HMW, and MMW adiponectin ($r = 0.22$; $r = 0.24$; and $r = 0.22$, respectively), and weak for sOB-R ($r = -0.03$), FABP4 ($r = 0.09$), resistin ($r = 0.11$), LMW adiponectin ($r = 0.06$), and IL6 ($r = 0.15$).

A.4.5.8 Results: Prediction of adipokine plasma concentrations

For the primary aim, we examined how much of the variance in the biomarker plasma concentrations can be explained by VAT mass as well as SAT mass, SAT gene expression and their interaction, in form of the total explained variance (adjusted R^2) from linear regression models (Supplementary Table A.38). The interaction of SAT mass and SAT gene expression can be interpreted as the relative metabolic activity of SAT projected onto its absolute mass. Overall, a substantial proportion of the plasma concentration was explained by these predictors for leptin and, albeit to a somewhat smaller extent, for FABP4. In contrast, the variance explained by these predictors was low for the remaining adipokines. In more detail, SAT gene expression explained 45%, SAT mass 72%, VAT mass 0%, and gene expression together with SAT 76% of variance in leptin plasma levels. For FABP4, the explained variance was 1% by gene expression, 31% by SAT, 1% by VAT, and 36% by SAT and gene expression. For sOB-R, the explained variance was 0% by gene expression, 13% by SAT, and 6% by VAT. For total, HMW, and MMW adiponectin, SAT mass as single predictor did not explain any of their variance, but VAT explained about 10-12% and gene expression explained about 6-7%. For IL-6, VAT explained 12%, SAT 1% and gene expression 1%. For resistin and LMW adiponectin, the variance explained by SAT, VAT, and gene expression was close to zero. The interaction of SAT gene expression and SAT mass accounted for 31% of leptin, and about 10% of sOB-R and FABP4 levels, but did not explain any variance on top of the main effects. When VAT was added to gene expression and SAT, the explained variance increased by 1% (leptin), 4% (sOB-R), 0% (resistin), 1% (FABP4), 8% (total adiponectin), 9% (HMW adiponectin), 8% (MMW adiponectin), 2% (LMW adiponectin), and 10% (IL6), see Supplementary Table A.38. It should be noted that without accounting for sex differences (cf. Supplementary Table A.35), the contribution of SAT and VAT cannot be assessed since they are sex-dependent, and otherwise they seem to explain 0% variance for some adipokines.

A.4.5.9 Results: Additional predictors of adipokine concentrations

In addition to the primary aim, we also investigated to what extent sex differences may explain variance in plasma adipokine concentrations (Supplementary Table A.39). While sex alone explained 28% of variance in leptin levels and 16% of FABP4 levels, it explained little of the variance in the remaining adipokines. Further, when sex was added to GE, SAT and VAT, it did not increase the explained variance in leptin levels, suggesting that variance in leptin differences between sexes were accounted for by differences in SAT mass and gene expression. Only for FABP4, sex explained additional variance (5%) on top of gene expression and AT

mass.

We also considered to what extent other factors may increase the explained variance in adipokine levels, including: 1) age, occupational training, physical activity, 2) all other anthropometric measures, 3) gene expression levels of all remaining adipokines, 4) plasma concentrations of all remaining adipokines, and 5) employment status, partner status, smoking, socioeconomic status and diabetes (Supplementary Table A.39). In general, small increases (at least 5%) in explained variance were only observed for the following adipokines: For sOB-R, the other anthropometric factors, leptin gene expression and leptin plasma concentration each explained about 6% additional variance (non-additively); for resistin, age, physical activity and FABP4 plasma concentration increased the explained variance from 1% to 14%; for FABP4, the largest increase was seen when other anthropometric factors (9%) or resistin plasma levels (6%) were included; for HMW adiponectin, the largest increases were observed when including further sociocultural factors (from 13% to 20%); for MMW adiponectin, when including all other gene expressions (from 12% to 25%); and for IL6 when including all other plasma levels (from 12% to 24%).

A.4.5.10 Additional results

Supplementary Tables A.40-A.43 show correlations (Pearson correlation coefficient r) between the mRNA expression of the different genes (Supplementary Table A.40), between the concentrations of the plasma concentrations (Supplementary Table A.41), between the different MRI and anthropometric measures (Supplementary Table A.42) and between the gene expression and plasma levels (Supplementary Table A.43). On the mRNA expression level, the highest correlations were observed between adiponectin and FABP4 ($r = 0.64$), adiponectin and leptin ($r = 0.37$), leptin and IL6 ($r = 0.34$). On the protein plasma level, the highest correlations were between leptin and FABP4 ($r = 0.53$) and leptin and leptin receptor ($r = -0.42$), beside the correlation between the different adiponectin molecular fractions. Regarding the obesity measures, BMI showed a high correlation with overall mass measures ($r = 0.85$ with TBV, $r = 0.80$ with TAT), intermediate correlation with SAT ($r = 0.69$), and only moderate or small correlation with VAT and CAT mass ($r < 0.5$). WC and WHR had higher correlation with internal AT ($r = 0.80/0.76$ with VAT, $r = 0.74/0.72$ with CAT), but very different correlations with overall mass measures ($r = 0.88/0.51$ with TBV; $r = 0.59/0.05$ with TAT; $r = 0.36/-0.22$ with SAT). Regarding the MRI measures, the correlation of TAT was high with SAT ($r = 0.95$) but only moderate with TBV ($r = 0.68$) and even smaller with VAT ($r = 0.42$).

A.4.5.11 Discussion

This study investigated to what extent plasma adipokine concentrations can be predicted based on MRI-derived AT mass of different body compartments, including SAT and VAT, and by SAT gene expression. Except for leptin and in part FABP4, surprisingly little variance was explained for the remaining adipokines. For all adipokines, AT mass was the main contributor to these percentages and played a much larger role compared to SAT gene expression. These data suggest that factors other than obesity and AT gene expression may be main determinants of these plasma adipokine concentrations. Further, while these findings

do not contradict a potential role of these adipokines in disease development, they suggest that these adipokines are unlikely to account for a large proportion of the association between adiposity and disease risk observed in epidemiological studies.

Leptin is a 16 kDa protein hormone encoded by the ob gene that is mostly produced by white adipose tissue (Zhang et al., 1994). It was originally discovered in 1994 as a long-term regulator of food intake and energy balance acting in the hypothalamus (Zhang et al., 1994). Today, it is well known that leptin also effects metabolism, energy homeostasis, neuroendocrine, and immune function (Ibrahim, 2010; Mantzoros et al., 2011; Friedman, 2016). In our analysis, a large proportion of the variance (72%) of circulating leptin levels was predicted by SAT. While gene expression explained 45% when considered alone, it contributed only little additional information when added to SAT mass. These data confirm that leptin is primarily a biomarker for SAT mass. However, SAT mass was also associated with leptin gene expression, so that the effect of SAT on circulating leptin could also contain part of the effect of leptin gene expression. The association of VAT mass with leptin levels was much smaller, but some of the remaining unexplained variance of leptin levels might be explained by VAT gene expression which was not assessed in this study.

FABP4 acts as a carrier for fatty acids and while it is also expressed in the liver or placenta, it is primarily expressed in adipocytes and macrophages (Furuhashi & Hotamisligil, 2008). In our study, about one third of the variance of FABP4 was explained, with rather similar contributions of SAT and VAT (Supplementary Table A.35). Gene expression did not contribute at all to the explained variance. Thus, somewhat similar as for leptin, these data suggest that FABP4 is a biomarker for AT mass, although about two thirds of the variance remain unexplained.

sOB-R is the main leptin-binding protein, and its expression has been reported in multiple other tissues besides AT, most notably in the hypothalamus (Zhang et al., 1994). The exact role of sOB-R is not entirely clear. Binding to sOB-R may delay the clearance of leptin from the circulation and thereby prolong bioavailability, but it may also neutralize the action of leptin and thereby reduce bioavailability (Zastrow et al., 2003; Zhang & Scarpace, 2009). In humans, sOB-R is inversely related to obesity and insulin resistance (Magni et al., 2005). Interestingly, recent studies found high sOB-R to be associated with lower risk of type 2 diabetes and colorectal cancer, independent of leptin concentrations (Sun et al., 2009; Aleksandrova et al., 2012a). In our analysis, the variance in plasma levels explained by SAT, VAT or gene expression was low (16%). Of these, SAT and VAT were the main contributors whereas gene expression did not contribute at all.

Adiponectin is involved, among others, in the regulation of insulin sensitivity and energy homeostasis. It circulates in different molecular fractions and is predominantly synthesized and secreted from adipocytes in VAT and SAT with different findings regarding their metabolic activity (Fredriksson et al., 2006; Goldstein et al., 2009; Pischon, 2009; Kovacova et al., 2012; Lihn et al., 2014). In our analysis, VAT explained slightly more to the variance of total and HMW adiponectin as compared to SAT, and was the main contributor for MMW adiponectin. However, the total variance explained was very low and only about 10-12% for total, HMW, and MMW adiponectin, and 2% for LMW adiponectin. This is surprising, given that adiponectin is often considered as ”the” classical adipokine. There is a vast

body of literature on the hormonal and metabolic effects of adiponectin, and on the potential role in disease development (Woodward et al., 2016). In contrast to most other adipokines, adiponectin is inversely associated with obesity, and it is considered to have beneficial effects: it increases insulin sensitivity, decreases inflammation, and it is inversely associated with risk of type 2 diabetes, coronary heart disease, and several types of cancer (Spranger et al., 2003; Pischon et al., 2004; Aleksandrova et al., 2012b). However, the effects of adiponectin are not as straightforward as initially considered, since studies have also found high adiponectin levels to be associated with adverse health outcome in certain populations, particularly in those with preexisting cardiovascular disease (Wannamethee et al., 2007; Dekker et al., 2008). It is not clear why the variance in plasma adiponectin levels explained by AT and gene expression observed in our study was so low. Measurement error appears unlikely since assays we used to measure adiponectin are well established and have been used in several other studies, the coefficients of variation were low (11% or less for total adiponectin and the fractions), and the strength of the correlation between adiponectin and BMI and waist circumference was similar to what has been reported in other studies (Maahs et al., 2007). These were also true for all other investigated adipokines, and the substantial correlations of AT measures with leptin can serve as a positive control. We speculate whether other factors, for example, post-translational modifications, may account for some of the variance in adiponectin levels.

Resistin is an adipokine that was initially reported to be elevated in obese C57Bl/6J mice on a high-fat diet and suppressed by treatment with thiazolidinediones (Steppan et al., 2001). Treatment of wild-type mice with recombinant resistin resulted in insulin resistance, whereas administration of an antiresistin antibody increased insulin sensitivity in obese and insulin-resistant animals (Steppan et al., 2001). However, the role of resistin in humans is less clear. Studies have found high plasma resistin levels to be associated with inflammation and with higher risk of coronary heart disease and with other inflammatory related diseases (Pischon et al., 2005; Weikert et al., 2008; Yang et al., 2016; Akram et al., 2017). However, in contrast to mice, human resistin is expressed at lower levels in adipocytes but at higher levels in circulating blood monocytes (Savage et al., 2001). This is in line with our observation that AT and gene expression did not predict circulating resistin levels. Taken together, these data question whether resistin can truly be considered an adipokine in humans.

Finally, the proinflammatory cytokine IL6 is secreted by multiple cell types including monocytes, macrophages, and adipocytes (Pischon, 2009). Although it was initially considered to be primarily an unspecific pro-inflammatory cytokine, studies have shown higher IL-6 levels in obese persons, and IL-6 was also shown to be directly secreted from the adipose tissue (Kern et al., 2001; Wallenius et al., 2002). This has led to the notion that obesity is associated with low-grade subclinical inflammation (Xu et al., 2003a). Nevertheless, results from our study suggest that obesity explains only a small proportion of circulating IL-6 levels in humans.

Without the assessment of VAT gene expression and secretion rates from SAT and VAT, parts of the overall molecular picture remain unclear. However, assessing VAT in a population-based study is rarely possible, and it might be questionable if VAT gene expression measured after bariatric or other surgeries allows for a valid approximation of the metabolic activity in healthy participants. Since adipokine secretion as well as their biological function can be

regulated on the transcriptional or post-transcriptional level, gene expression which might not be the best proxy for the metabolic activity of AT and might differ from the secretion rate (Liu & Liu, 2012; Hotamisligil & Bernlohr, 2015). Also, some adipokines are expressed and secreted into plasma from other tissues (Jamaluddin et al., 2012; Hunter & Jones, 2015). Furthermore, some factors such as lifestyle factors, clinical parameters and other circulating proteins which that have been identified in previous studies to affect adipokine serum levels were not included in the present study (Thomas et al., 2000; Chan et al., 2002; Norata et al., 2007; Yamagishi et al., 2009; Hong et al., 2015; Ishimura et al., 2013). The additional lifestyle factors and other predictors investigated in the current study did not explain a substantial amount of the unexplained variance. Interesting markers for future studies include diet measures (Campbell et al., 2013) and genetic markers that don't affect mRNA abundance but circulating levels through other processes, which could also help to infer more about the causal mechanisms and molecular processes through mendelian randomization techniques (Smith & Ebrahim 2003).

In conclusion, our study shows that while for leptin, most of the variance in plasma concentrations can be explained by AT mass (particularly SAT mass) and SAT gene expression, this is less so for FABP4. In contrast, and counterintuitively, most of the variance in the plasma concentrations of the so-called adipokines sOB-R, resistin, adiponectin, and IL6 cannot be explained by AT mass or SAT gene expression. These data suggest that other factors are the main determinants of these plasma concentrations. Further, while these findings do not contradict a potential role of these adipokines in disease development, they suggest that these adipokines are unlikely to account for a large proportion of the association between adiposity and disease risk observed in epidemiological studies.

A.5 Supplementary Tables

A.5.1 List of supplementary tables

A.5.2 Simulation study: single-marker versus multi-marker tests

Error Distribution	SKAT	SKAT-O	Burden	SMT - BH	SMT - Bonferroni	SMT - Bonferroni MAC>1	SMT - Bonferroni MAC>2
$N(0,1)$	4.94×10^{-2}	5.22×10^{-2}	5.01×10^{-2}	4.39×10^{-2}	4.93×10^{-2}	–	–
t_8	5.17×10^{-2}	5.39×10^{-2}	5.00×10^{-2}	12.68×10^{-2}	12.01×10^{-2}	6.53×10^{-2}	5.09×10^{-2}
t_4	6.13×10^{-2}	6.09×10^{-2}	5.00×10^{-2}	21.30×10^{-2}	20.61×10^{-2}	10.64×10^{-2}	7.57×10^{-2}
$\log N(0,1)$	8.84×10^{-2}	7.95×10^{-2}	4.96×10^{-2}	33.63×10^{-2}	33.05×10^{-2}	18.88×10^{-2}	13.47×10^{-2}

Supplementary Table A.1: Empirical type I error estimates of the SMT and MMTs for the nominal level $\alpha = 0.05$ when the distribution of the phenotype is misspecified. Data was generated from the null model with size $n = 1,000$ for $m = 10,000,000$ replicates, with the error distribution of the phenotype given covariates chosen to be from a standard normal (cf. Table 4.1), t_4, t_8, or log-standard normal distribution. Adjustments for multiple testing of all SNVs in a gene with the SMT were done using the BH and Bonferroni corrections. The type I error estimates are based on analyses using all rare and (non-causal) common SNVs in a gene, as well as restricting the analysis to SNVs with at least 2 ('MAC>1') or 3 ('MAC>2') observed minor alleles for the SMT. For other nominal levels (e.g., $10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 2.5 \times 10^{-6}$, SKAT, SKAT-O and the SMT showed similar results and inflated type I errors, with a much higher inflation for the SMT.

Scenario	Using all rare and common SNVs				Using all rare SNVs			
	SKAT	SKAT-O	Burden	SMT	SKAT	SKAT-O	Burden	SMT
1	0.341	0.337	0.107	0.546	0.358	0.357	0.242	0.563
2	0.199	0.187	0.080	0.192	0.209	0.204	0.140	0.200
3	0.127	0.124	0.062	0.100	0.130	0.130	0.099	0.106
4	0.532	0.534	0.170	0.736	0.553	0.559	0.414	0.752
5	0.299	0.293	0.102	0.282	0.316	0.317	0.230	0.297
6	0.168	0.170	0.076	0.130	0.182	0.185	0.148	0.137
7	0.764	0.783	0.290	0.907	0.792	0.815	0.698	0.917
8	0.477	0.492	0.159	0.440	0.507	0.532	0.440	0.458
9	0.289	0.311	0.115	0.209	0.308	0.337	0.284	0.220
10	0.986	0.996	0.583	0.997	0.993	0.999	0.992	0.998
11	0.825	0.895	0.352	0.730	0.861	0.928	0.896	0.755
12	0.580	0.695	0.242	0.407	0.619	0.752	0.721	0.433
13	0.352	0.346	0.113	0.554	0.368	0.361	0.245	0.570
14	0.188	0.180	0.075	0.187	0.197	0.193	0.133	0.196
15	0.113	0.108	0.060	0.094	0.119	0.113	0.089	0.099
16	0.516	0.497	0.136	0.727	0.540	0.522	0.310	0.744
17	0.300	0.277	0.086	0.280	0.318	0.296	0.170	0.293
18	0.165	0.151	0.070	0.132	0.177	0.163	0.107	0.139
19	0.761	0.747	0.227	0.903	0.785	0.777	0.523	0.913
20	0.473	0.454	0.131	0.432	0.503	0.489	0.310	0.456
21	0.282	0.269	0.093	0.198	0.297	0.290	0.196	0.211
22	0.981	0.981	0.410	0.996	0.988	0.989	0.837	0.997
23	0.806	0.811	0.229	0.713	0.842	0.849	0.625	0.737
24	0.541	0.549	0.163	0.374	0.583	0.602	0.445	0.399
25	0.352	0.335	0.098	0.543	0.367	0.351	0.200	0.561
26	0.195	0.179	0.075	0.188	0.206	0.188	0.117	0.199
27	0.125	0.114	0.058	0.093	0.128	0.117	0.078	0.098
28	0.528	0.499	0.132	0.732	0.548	0.519	0.274	0.749
29	0.295	0.270	0.084	0.286	0.314	0.292	0.159	0.303
30	0.163	0.148	0.063	0.125	0.172	0.158	0.105	0.134
31	0.755	0.727	0.187	0.903	0.782	0.752	0.390	0.914
32	0.460	0.422	0.103	0.416	0.488	0.452	0.224	0.438
33	0.270	0.235	0.076	0.184	0.289	0.255	0.136	0.200
34	0.981	0.973	0.280	0.995	0.986	0.981	0.577	0.997
35	0.801	0.763	0.156	0.692	0.838	0.802	0.346	0.718
36	0.526	0.472	0.111	0.352	0.566	0.514	0.233	0.376

Supplementary Table A.2: Power estimates of the SMT and MMTs under the nominal α level of 0.05. Data was generated under the alternative-hypothesis model described in scenarios 1-36 in Table 4.1 with size $n = 1,000$ for $m = 10,000$ replicates. Adjustments for multiple testing of all SNVs in a gene with the SMT were done using the BH correction. The power results are provided for analyses using all rare and (non-causal) common SNVs in a gene, and for using all rare SNVs in a gene by excluding the common SNVs from the analysis.

Scenario	Using all rare and common SNVs				Using all rare SNVs			
	SKAT	SKAT-O	Burden	SMT	SKAT	SKAT-O	Burden	SMT
1	0.118	0.111	0.012	0.156	0.129	0.122	0.050	0.159
2	0.007	0.008	0.002	0.005	0.008	0.009	0.004	0.005
3	0.001	0.001	<0.001	0.001	0.001	0.001	0.001	0.001
4	0.223	0.215	0.024	0.263	0.243	0.235	0.121	0.268
5	0.021	0.021	0.004	0.014	0.022	0.022	0.015	0.014
6	0.004	0.004	<0.001	0.002	0.003	0.004	0.002	0.002
7	0.410	0.414	0.065	0.422	0.443	0.451	0.306	0.428
8	0.063	0.071	0.012	0.043	0.067	0.077	0.056	0.045
9	0.014	0.016	0.003	0.009	0.015	0.018	0.014	0.009
10	0.797	0.867	0.207	0.698	0.849	0.927	0.874	0.707
11	0.248	0.350	0.070	0.155	0.273	0.421	0.388	0.156
12	0.093	0.128	0.036	0.059	0.095	0.147	0.139	0.060
13	0.124	0.117	0.013	0.158	0.134	0.128	0.053	0.162
14	0.007	0.006	0.001	0.005	0.007	0.007	0.003	0.005
15	0.001	0.001	0	0.001	0.001	0.001	<0.001	0.001
16	0.210	0.200	0.016	0.250	0.227	0.216	0.068	0.256
17	0.015	0.014	0.001	0.008	0.016	0.015	0.006	0.008
18	0.002	0.002	<0.001	0.001	0.002	0.002	0.001	0.001
19	0.393	0.380	0.038	0.406	0.425	0.415	0.178	0.414
20	0.050	0.048	0.007	0.029	0.055	0.054	0.028	0.031
21	0.011	0.015	0.001	0.006	0.011	0.012	0.007	0.006
22	0.764	0.771	0.108	0.667	0.815	0.828	0.522	0.678
23	0.203	0.212	0.038	0.110	0.228	0.246	0.159	0.111
24	0.057	0.063	0.015	0.035	0.060	0.067	0.053	0.035
25	0.123	0.114	0.009	0.161	0.137	0.128	0.035	0.164
26	0.007	0.006	<0.001	0.004	0.008	0.007	0.002	0.005
27	<0.001	<0.001	0	<0.001	0.001	<0.001	<0.001	<0.001
28	0.212	0.199	0.015	0.252	0.231	0.217	0.060	0.257
29	0.016	0.014	0.001	0.009	0.018	0.015	0.005	0.009
30	0.002	0.001	<0.001	0.001	0.001	0.001	0.001	0.001
31	0.393	0.375	0.025	0.403	0.424	0.405	0.105	0.411
32	0.039	0.034	0.004	0.020	0.045	0.038	0.013	0.021
33	0.006	0.006	0	0.003	0.006	0.006	0.003	0.003
34	0.746	0.732	0.047	0.661	0.792	0.782	0.207	0.671
35	0.166	0.147	0.009	0.068	0.187	0.165	0.042	0.070
36	0.035	0.031	0.004	0.016	0.038	0.033	0.012	0.017

Supplementary Table A.3: Power estimates of the SMT and MMTs under the nominal α level of 2.5×10^{-6}. Data was generated under the alternative-hypothesis model described in scenarios 1-36 in Table 4.1 with size $n = 1,000$ for $m = 10,000$ replicates. Adjustments for multiple testing of all SNVs in a gene with the SMT were done using the BH correction. Power results are provided for analyses using all rare and (non-causal) common SNVs in a gene, and for using all rare SNVs in a gene by excluding the common SNVs from the analysis.

Scenario	Using all rare and common SNVs				Using all rare SNVs			
	SKAT	SKAT-O	Burden	SMT	SKAT	SKAT-O	Burden	SMT
1	0.276	0.271	0.063	0.558	0.282	0.280	0.173	0.561
2	0.153	0.148	0.027	0.195	0.160	0.155	0.077	0.196
3	0.060	0.058	0.011	0.047	0.062	0.061	0.029	0.048
4	0.462	0.463	0.117	0.774	0.473	0.480	0.341	0.777
5	0.269	0.265	0.053	0.315	0.279	0.276	0.162	0.318
6	0.117	0.116	0.024	0.088	0.124	0.123	0.075	0.090
7	0.722	0.764	0.243	0.934	0.737	0.792	0.677	0.936
8	0.484	0.497	0.109	0.517	0.505	0.524	0.387	0.521
9	0.252	0.266	0.063	0.167	0.269	0.290	0.208	0.170
10	0.984	0.999	0.583	0.999	0.990	1	0.997	1
11	0.862	0.938	0.309	0.813	0.883	0.965	0.931	0.816
12	0.632	0.765	0.178	0.374	0.664	0.827	0.771	0.379

Supplementary Table A.4: Power estimates of the SMT and MMTs under the nominal α level of 2.5×10^{-6}, for a sample size of 5,000 and genes including 58 SNVs on average. Data was generated under the alternative-hypothesis model described in scenarios 1-12 in Table 4.1 with size $n = 5,000$ for $m = 10,000$ replicates. Adjustments for multiple testing of all SNVs in a gene with the SMT were done using the BH correction. Power results are provided for analyses using all rare and (non-causal) common SNVs in a gene, and for using all rare SNVs in a gene by excluding the common SNVs from the analysis.

Scenario	Using all rare and common SNVs				Using all rare SNVs			
	SKAT	SKAT-O	Burden	SMT	SKAT	SKAT-O	Burden	SMT
1	0.861	0.865	0.304	0.998	0.872	0.876	0.724	0.998
2	0.622	0.615	0.191	0.782	0.664	0.660	0.500	0.784
3	0.387	0.391	0.133	0.335	0.414	0.425	0.329	0.339
4	0.989	0.992	0.468	1	0.992	0.995	0.963	1
5	0.887	0.895	0.304	0.936	0.915	0.930	0.854	0.937
6	0.694	0.717	0.241	0.605	0.737	0.775	0.692	0.608
7	1	1	0.681	1	1	1	1	1
8	0.994	0.998	0.480	0.994	0.997	1	0.997	0.994
9	0.944	0.971	0.364	0.874	0.971	0.991	0.982	0.874
10	1	1	0.877	1	1	1	1	1
11	1	1	0.788	1	1	1	1	1
12	1	1	0.679	0.998	1	1	1	0.998

Supplementary Table A.5: Power estimates of the SMT and MMTs under the nominal α level of 2.5×10^{-6}, for a sample size of 5,000 and genes including 572 SNVs on average. Data was generated under the alternative-hypothesis model described in scenarios 1-12 in Table 4.1 with size $n = 5,000$ for $m = 10,000$ replicates. Adjustments for multiple testing of all SNVs in a gene with the SMT were done using the BH correction. Power results are provided for analyses using all rare and (non-causal) common SNVs in a gene, and for using all rare SNVs in a gene by excluding the common SNVs from the analysis.

Scenario	SKAT	SKAT-O	Burden	SMT
1	0.363	0.350	0.002	0.515
4	0.565	0.554	0.006	0.715
7	0.860	0.864	0.041	0.928
8	0.388	0.406	0.002	0.190
12	0.531	0.733	0.002	0.231

Supplementary Table A.6: Power estimates of the SMT and MMTs under the nominal α level of 2.5×10^{-6}, for a sample size of 2,000 and genes including 67 SNVs. As genetic data, the available genotypes of the 67 SNVs of $n = 2,000$ individuals in the 'SKAT.example' data in the `SKAT` R package were used in each replicate. Data was generated under the alternative-hypothesis model described in scenarios 1, 4, 7, 8, 12 in Table 4.1 with size $n = 2,000$ for $m = 10,000$ replicates. Adjustments for multiple testing of all SNVs in a gene with the SMT were done using the BH correction. Power results are provided for analyses using all rare and (non-causal) common SNVs in a gene.

Scenario	SKAT	SKAT-O	Burden	SMT - BH	SMT - Bonferroni
1	0.118	0.111	0.012	0.156	0.153
2	0.007	0.008	0.002	0.005	0.004
3	0.001	0.001	<0.001	0.001	<0.001
4	0.223	0.215	0.024	0.263	0.259
5	0.021	0.021	0.004	0.014	0.011
6	0.004	0.004	<0.001	0.002	0.002
7	0.410	0.414	0.065	0.422	0.415
8	0.063	0.071	0.012	0.043	0.040
9	0.014	0.016	0.003	0.009	0.007
10	0.797	0.867	0.207	0.698	0.692
11	0.248	0.350	0.070	0.155	0.147
12	0.093	0.128	0.036	0.059	0.053

Supplementary Table A.7: Power estimates of the SMT and MMTs under the nominal α level of 2.5×10^{-6} to compare the BH and Bonferroni correction for SMTs. Data was generated under the alternative-hypothesis model described in scenarios 1-12 in Table 4.1 with size $n = 1,000$ for $m = 10,000$ replicates. Adjustments for multiple testing of all SNVs in a gene with the SMT were done using the BH or the Bonferroni correction. Power results are provided for analyses using all rare and (non-causal) common SNVs in a gene.

α	Scenario	SKAT	SKAT-O	Burden	SMT
0.05	1	0.166	0.157	0.070	0.062
	4	0.247	0.238	0.086	0.097
	7	0.397	0.408	0.139	0.162
	8	0.164	0.178	0.083	0.043
	12	0.200	0.270	0.112	0.051
2.5×10^{-6}	1	0.003	0.003	<0.001	0
	4	0.009	0.009	0.001	0
	7	0.029	0.032	0.005	<0.001
	8	0.002	0.002	<0.001	0
	12	0.003	0.005	<0.001	0

Supplementary Table A.8: Power estimates of the SMT and MMTs for analyzing a binary trait under the nominal α levels of 0.05 and 2.5×10^{-6}. Case-control data was generated under the alternative-hypothesis model described in scenarios 1, 4, 7, 8, 12 in Table 4.1 with size $n = 1,000$ for $m = 10,000$ replicates for a binary trait. Adjustments for multiple testing of all SNVs in a gene with the SMT were done using the BH correction. All rare and (non-causal) common SNVs were incorporated in the analysis.

Scenario	Gene-level		SNV-level	
	SKAT-O	SMT	SMT for SNVs with MAF = 0.002	SMT for SNVs with MAF = 0.005
13	0.117	0.158	0.022	0.188
14	0.006	0.005	0	0
15	0.001	0.001	0	0.010
16	0.200	0.250	0.016	0.220
17	0.014	0.008	0	0.006
18	0.002	0.001	0	0
19	0.380	0.406	0.022	0.203
20	0.048	0.029	0	0.034
21	0.015	0.006	0	0.005
22	0.771	0.667	0.023	0.225
23	0.212	0.110	<0.001	0.055
24	0.063	0.035	0	0.010

Supplementary Table A.9: Comparison of power estimates for identifying a causal gene and a causal SNV with a given MAF. Data was generated under the alternative-hypothesis model described in scenarios 13-24 in Table 4.1 with size $n = 1,000$ for $m = 10,000$ replicates. The nominal α was set to 2.5×10^{-6} for the gene-level evaluation of SMT and SKAT-O (representing the Bonferroni-correction for testing 20,000 genes), and to 10^{-7} for the SNV-level evaluation of SMT (representing the Bonferroni-correction for testing 500,000 SNVs). Adjustments for multiple testing of all SNVs in a gene were done using the BH-correction. The power of the SMT for identifying a causal SNV with a given MAF is based on all causal SNVs in the $m = 10,000$ replicates with the specified MAF. It is estimated by the number of significant causal SNVs (with p-values smaller than 10^{-7} divided by the total number of causal SNVs with the specified MAF. The power for testing singletons and doubletons is even smaller than the power of testing SNVs with MAF=0.002.

Scenario	SNV with MAF = 0.0005	SNV with MAF = 0.001	SNV with MAF = 0.002	SNV with MAF = 0.005
13	0.133	0.280	0.555	0.875
14	0.020	0.033	0.055	0.200
15	0.008	0.011	0.022	0.039
16	0.131	0.283	0.542	0.873
17	0.015	0.032	0.072	0.179
18	0.009	0.008	0.022	0.052
19	0.127	0.268	0.506	0.859
20	0.018	0.033	0.069	0.290
21	0.008	0.010	0.021	0.082
22	0.114	0.237	0.461	0.801
23	0.021	0.037	0.072	0.267
24	0.009	0.014	0.025	0.158

Supplementary Table A.10: Conditional power estimates of the SMT for identifying a causal SNV, given that the gene contains a causal SNV. Data was generated under an alternative-hypothesis model described in scenarios 13-24 in Table 4.1 with size $n = 1,000$ for $m = 10,000$ replicates. The nominal α level was set to 0.002, representing the Bonferroni-correction for testing 25 SNVs in a gene (assuming 500,000 SNVs in total within 20,000 genes). The power of the SMT for identifying a causal SNV with a given MAF is based on all causal SNVs in the $m = 10,000$ replicates with the specified MAF. It is estimated by the number of significant causal SNVs (with p-values smaller than 0.002) divided by the total number of causal SNVs with the specified MAF.

Gene	SKAT		SKAT-O		Burden		SMT	
	All SNVs	Rare SNVs	All SNVs	Rare SNVs	All SNVs	Rare SNVs	All SNVs	Rare SNVs
INSR	0.51	0.73	0.72	0.76	0.27	0.70	0.14	0.07
RRAS	0.33	0.32	0.49	0.46	0.59	0.34	0.26	0.18
ZNF101	0.94	0.93	0.91	0.87	0.97	0.72	0.98	0.97
ELAVL3	0.46	0.44	0.29	0.42	0.12	0.27	0.76	0.74
RGL3	0.21	0.13	0.32	0.22	0.46	0.22	0.03	0.02
AMH	0.30	0.24	0.46	0.38	0.65	0.73	0.18	0.12
DOT1L	0.59	0.60	0.78	0.79	0.43	0.73	0.70	0.51
PLEKHJ1	0.59	0.50	0.79	0.72	0.33	0.67	0.18	0.13
SF3A2	0.50	0.36	0.70	0.54	0.41	0.97	0.23	0.15

Supplementary Table A.11: Gene-level p-values for the association tests of candidate genes with SBP in the Genetic Analysis 19 data analysis: Unadjusted p-values from gene-level genetic association analysis with SBP, of all common and rare SNVs ('all SNVs'), and only rare SNVs ('rare SNVs') in 9 candidate genes. Adjustments for multiple testing of all SNVs in SMTs in a gene were done using the BH-correction.

A.5.3 Simulation study: C-JAMP

Dependence	MAC	$\overline{\widehat{\beta}_{XY}}$	$SD(\widehat{\beta}_{XY})$	$\overline{\widehat{SE}(\widehat{\beta}_{XY})}$
$\tau = 0.2$	1	-0.048	1.00	0.99
	2	-0.012	0.70	0.69
	4	-0.002	0.49	0.49
	10	0.002	0.31	0.31
$\tau = 0.5$	1	0.308	1.00	1.20
	2	0.112	0.67	0.69
	4	0.046	0.46	0.46
	10	0.017	0.28	0.28
$\tau = 0.8$	1	0.610	1.00	0.90
	2	0.213	0.62	0.57
	4	0.079	0.41	0.39
	10	0.027	0.25	0.24

Supplementary Table A.12: Bias of the point estimates and standard error estimates of C-JAMP Wald test statistics. Data was generated from the null model described in scenario 0 in Table 4.4 for $n = 1,000$ individuals with $m = 100,000$ replicates, and under different dependencies between the two traits Y_1, Y_2 (Kendall's $\tau = 0.2, 0.5, 0.8$). Shown are the mean of the coefficient estimates, standard deviation of the coefficient estimates, and mean standard error estimates of the genetic effect based on all SNVs with the given MAC in all replicates.

Nominal α	Dependence	C-JAMP	Nominal α	Dependence	C-JAMP
	$\tau = 0.1$	0.051		$\tau = 0.1$	1.5×10^{-4}
	$\tau = 0.2$	0.050		$\tau = 0.2$	1.8×10^{-4}
	$\tau = 0.3$	0.044		$\tau = 0.3$	1.0×10^{-4}
	$\tau = 0.4$	0.033		$\tau = 0.4$	6.0×10^{-5}
5×10^{-2}	$\tau = 0.5$	0.031	10^{-4}	$\tau = 0.5$	1.3×10^{-4}
	$\tau = 0.6$	0.068		$\tau = 0.6$	3.1×10^{-4}
	$\tau = 0.7$	0.161		$\tau = 0.7$	1.6×10^{-3}
	$\tau = 0.8$	0.206		$\tau = 0.8$	3.1×10^{-3}
	$\tau = 0.9$	0.234		$\tau = 0.9$	1.9×10^{-2}
	$\tau = 0.1$	0.011		$\tau = 0.1$	2.0×10^{-5}
	$\tau = 0.2$	0.011		$\tau = 0.2$	3.1×10^{-5}
	$\tau = 0.3$	0.010		$\tau = 0.3$	2.0×10^{-5}
	$\tau = 0.4$	0.007		$\tau = 0.4$	1.0×10^{-5}
10^{-2}	$\tau = 0.5$	0.007	10^{-5}	$\tau = 0.5$	0×10^{-5}
	$\tau = 0.6$	0.016		$\tau = 0.6$	6.0×10^{-5}
	$\tau = 0.7$	0.051		$\tau = 0.7$	3.1×10^{-4}
	$\tau = 0.8$	0.072		$\tau = 0.8$	6.8×10^{-4}
	$\tau = 0.9$	0.103		$\tau = 0.9$	1.1×10^{-2}
	$\tau = 0.1$	1.3×10^{-3}			
	$\tau = 0.2$	1.2×10^{-3}			
	$\tau = 0.3$	1.1×10^{-3}			
	$\tau = 0.4$	8.5×10^{-4}			
10^{-3}	$\tau = 0.5$	9.5×10^{-4}			
	$\tau = 0.6$	2.1×10^{-3}			
	$\tau = 0.7$	9.6×10^{-3}			
	$\tau = 0.8$	1.5×10^{-2}			
	$\tau = 0.9$	3.9×10^{-2}			

Supplementary Table A.13: Gene-level type I error estimates using unadjusted C-JAMP Wald test statistics. Data was generated from the null model described in scenario 0 in Table 4.4 for $n = 1,000$ individuals with $m = 100,000$ replicates, and under different dependencies between the two traits Y_1, Y_2 (Kendall's $\tau = 0.2, 0.5, 0.8$). Adjustments for multiple testing of all SNVs in a gene were done using the Bonferroni correction. Type I error estimates are for testing the association with the first trait, Y_1.

Dependence	all variants	MAF=0.0005	MAF=0.001	MAF=0.002	MAF=0.005
$\tau = 0.1$	0.053	0.055	0.052	0.052	0.052
$\tau = 0.2$	0.052	0.054	0.053	0.051	0.051
$\tau = 0.3$	0.049	0.046	0.051	0.052	0.053
$\tau = 0.4$	0.042	0.030	0.048	0.051	0.055
$\tau = 0.5$	0.040	0.025	0.046	0.051	0.054
$\tau = 0.6$	0.060	0.069	0.059	0.054	0.050
$\tau = 0.7$	0.084	0.120	0.075	0.059	0.053
$\tau = 0.8$	0.092	0.134	0.086	0.063	0.056
$\tau = 0.9$	0.094	0.138	0.089	0.065	0.055

Supplementary Table A.14: SNV-level type I error estimates using unadjusted C-JAMP Wald test statistics under different dependencies between the traits. SNV-level type I errors are estimated by the number of significant SNVs (with p-values smaller than 0.05) divided by the total number of tested SNVs. Data was generated from the null model described in scenario 0 in Table 4.4 for $n = 1,000$ individuals with $m = 100,000$ replicates. The nominal α level was set to 0.05. Type I error estimates are for testing the association with the first trait, Y_1. MAFs of 0.0005, 0.001, 0.002, 0.005 equal 1, 2, 4, and 10 copies of the minor allele, respectively.

Dependence	Estimate	Parameters	Values of Parameters
$\tau = 0.5$	Intercept	A	-2.25244
		B	11.04387
		C	0.62228
	Slope	A	-0.67510
		B	2.15985
		C	4.22380
$\tau = 0.8$	Intercept	A	-2.91350
		B	5.40099
		C	0.54384
	Slope	A	23.36902
		B	38.66929
		C	0.58071

Supplementary Table A.15: Parameter estimates for the functional form of the relationship between MAC and intercept/slope estimates for the C-JAMP Wald test statistic adjustment in equation (A.4.2), for different Kendall's τ.

Dependence	all variants	MAF=0.0005	MAF=0.001	MAF=0.002	MAF=0.005
$\tau = 0.2$	0.052	0.054	0.053	0.051	0.051
$\tau = 0.5$	0.049	0.049	0.044	0.053	0.053
$\tau = 0.8$	0.050	0.050	0.045	0.067	0.040

Supplementary Table A.16: SNV-based type I error estimates of C-JAMP using adjusted Wald test statistics for testing the association with SNVs of a specific MAF. SNV-level type I errors are estimated by the number of significant SNVs (with p-values smaller than 0.05) divided by the total number of tested SNVs with the specified MAF. Data was generated from the null model described in scenario 0 in Table 4.4 for $n = 1,000$ individuals with $m = 100,000$ replicates. The nominal α level was set to 0.05. Type I error estimates are for testing the association with the first trait, Y_1. MAFs of 0.0005, 0.001, 0.002, 0.005 equal 1, 2, 4, and 10 copies of the minor allele.

Dependence	Bonferroni correction	BH correction
$\tau = 0.2$	0.054	0.056
$\tau = 0.5$	0.053	0.064
$\tau = 0.8$	0.029	0.037

Supplementary Table A.17: Type I error estimates of C-JAMP using adjusted Wald test statistics under a misspecified joint model. Data was generated from a bivariate normal model under the null model described in scenario 0 in Table 4.4 for $n = 1,000$ individuals with $m = 1,000$ replicates, as described in section A.4.2.2. The nominal α level was set to 0.05. Adjustments for multiple testing of all SNVs in a gene were done using the Bonferroni or the BH correction. Type I error estimates are for testing the association with the first trait, Y_1.

Scenario	Dependence	Bonferroni correction	BH correction
13	$\tau = 0.2$	0.036	0.042
	$\tau = 0.5$	0.042	0.047
	$\tau = 0.8$	0.128	0.139
14	$\tau = 0.2$	0.056	0.063
	$\tau = 0.5$	0.039	0.041
	$\tau = 0.8$	0.078	0.092
15	$\tau = 0.2$	0.038	0.045
	$\tau = 0.5$	0.046	0.050
	$\tau = 0.8$	0.070	0.079

Supplementary Table A.18: Type I error estimates of C-JAMP for testing the association of SNVs with the second, not-associated trait in scenarios 13-15. Data was generated under the null model for the second trait Y_2 and under the alternative-hypothesis model for the first trait Y_1 as described in scenarios 13-15 in Table 4.4 for $n = 1,000$ individuals with $m = 1,000$ replicates. The nominal α level was set to 0.05. Adjustments for multiple testing of all SNVs in a gene were done using the Bonferroni or the BH correction. Type I error estimates are for testing the association with the second trait, Y_2.

α	Scenario	SKAT	SKAT-O	Burden	SMT	C-JAMP $\tau = 0.2$	C-JAMP $\tau = 0.5$	C-JAMP $\tau = 0.8$
0.05	1	0.341	0.337	0.107	0.546	0.560	0.582	0.627
	2	0.199	0.187	0.080	0.192	0.193	0.227	0.276
	3	0.127	0.124	0.062	0.100	0.098	0.106	0.134
	4	0.532	0.534	0.170	0.736	0.736	0.756	0.799
	5	0.299	0.293	0.102	0.282	0.285	0.342	0.404
	6	0.168	0.170	0.076	0.130	0.132	0.145	0.205
	7	0.764	0.783	0.290	0.907	0.904	0.923	0.942
	8	0.477	0.492	0.159	0.440	0.440	0.507	0.605
	9	0.289	0.311	0.115	0.209	0.203	0.242	0.303
	10	0.986	0.996	0.583	0.997	0.994	0.997	0.999
	11	0.825	0.895	0.352	0.730	0.724	0.798	0.865
	12	0.580	0.695	0.242	0.407	0.409	0.463	0.569
2.5×10^{-6}	1	0.118	0.111	0.012	0.156	0.175	0.220	0.284
	2	0.007	0.008	0.002	0.005	0.004	0.012	0.022
	3	0.001	0.001	<0.001	0.001	0.001	0.001	0.002
	4	0.223	0.215	0.024	0.263	0.285	0.348	0.418
	5	0.021	0.021	0.004	0.014	0.016	0.023	0.045
	6	0.004	0.004	<0.001	0.002	0.002	0.004	0.009
	7	0.410	0.414	0.065	0.422	0.453	0.540	0.630
	8	0.063	0.071	0.012	0.043	0.042	0.059	0.091
	9	0.014	0.016	0.003	0.009	0.009	0.013	0.024
	10	0.797	0.867	0.207	0.698	0.723	0.820	0.888
	11	0.248	0.350	0.070	0.155	0.160	0.192	0.257
	12	0.093	0.128	0.036	0.059	0.056	0.077	0.110

Supplementary Table A.19: Power estimates of C-JAMP in comparison to the univariate SMT and MMTs under the nominal α levels of 0.05 and 2.5×10^{-6}. Data was generated under an alternative-hypothesis model described in scenarios 1-12 in Table 4.4 for $n = 1,000$ individuals with $m = 10,000$ replicates. Adjustments for multiple testing of all SNVs in a gene were done using the Bonferroni correction for C-JAMP and the BH-correction for the SMT. Results of C-JAMP are based on adjusted Wald test statistics. Power estimates are for testing the association with the first trait, Y_1. These results are also shown in Figure 4.3.

τ	Scenario	C-JAMP	aSPUset	aSPUset-Score
0.2	1	0.560	0.127	0.375
	2	0.193	0.080	0.115
	3	0.098	0.060	0.077
	4	0.736	0.191	0.612
	5	0.285	0.111	0.182
	6	0.132	0.077	0.098
	7	0.904	0.307	0.894
	8	0.440	0.155	0.351
	9	0.203	0.110	0.160
	10	0.994	0.600	1
	11	0.724	0.348	0.801
	12	0.409	0.229	0.424
0.5	1	0.582	0.110	0.284
	2	0.227	0.071	0.089
	3	0.106	0.065	0.068
	4	0.756	0.154	0.485
	5	0.342	0.096	0.131
	6	0.145	0.069	0.077
	7	0.923	0.258	0.799
	8	0.507	0.136	0.247
	9	0.242	0.100	0.120
	10	0.997	0.534	1
	11	0.798	0.297	0.641
	12	0.463	0.198	0.310
0.8	1	0.627	0.096	0.172
	2	0.276	0.065	0.076
	3	0.134	0.057	0.062
	4	0.799	0.150	0.303
	5	0.404	0.085	0.098
	6	0.205	0.068	0.072
	7	0.942	0.233	0.576
	8	0.605	0.127	0.167
	9	0.303	0.095	0.101
	10	0.999	0.502	0.963
	11	0.865	0.279	0.422
	12	0.569	0.180	0.220

Supplementary Table A.20: Power estimates of C-JAMP in comparison to the multivariate tests under the nominal α level of 0.05, for different dependences τ between Y_1 and Y_2. Data was generated under an alternative-hypothesis model described in scenarios 1-12 in Table 4.4 for $n = 1,000$ individuals with $m = 10,000$ replicates. Adjustments for multiple testing of all SNVs in a gene were done using the Bonferroni correction for C-JAMP. Results of C-JAMP are based on adjusted Wald test statistics. Power estimates are for testing the association with the first trait, Y_1 for C-JAMP, and with respect to both traits for aSPUset and aSPUset-Score.

τ	α	Scenario	% causal variants	c_{Y_1}	c_{Y_2}	C-JAMP	aSPUset	aSPUset-Score
0.2	0.05	13	10%	0.6	0	0.763	0.183	0.495
		16	10%	0.6	0.1	0.738	0.171	0.473
		17	10%	0.6	0.2	0.740	0.170	0.456
		4	10%	0.6	0.6	0.736	0.191	0.612
		14	20%	0.3	0	0.450	0.149	0.265
		8	20%	0.3	0.3	0.440	0.155	0.351
		15	50%	0.2	0	0.430	0.212	0.324
		12	50%	0.2	0.2	0.409	0.229	0.424
	2.5×10^{-6}	13	10%	0.6	0	0.290	-	-
		16	10%	0.6	0.1	0.273	-	-
		17	10%	0.6	0.2	0.294	-	-
		4	10%	0.6	0.6	0.285	-	-
		14	20%	0.3	0	0.039	-	-
		8	20%	0.3	0.3	0.042	-	-
		15	50%	0.2	0	0.060	-	-
		12	50%	0.2	0.2	0.056	-	-

Supplementary Table A.21: Power estimates of C-JAMP and the multivariate tests for nominal α levels of 0.05 and 2.5×10^{-6}. Data was generated under scenarios 13-17 in Table 4.4 for $n = 1,000$ individuals with $m = 1,000$ replicates, and scenarios 4, 8, 12 from Supplementary Tables A.20 are shown for comparison. Adjustments for multiple testing of all SNVs in a gene with C-JAMP were done using the Bonferroni correction. Results of C-JAMP are based on adjusted Wald test statistics. Power estimates are for testing the association with the first trait, Y_1 for C-JAMP, and with respect to both traits for aSPUset and aSPUset-Score.

A.5.4 Simulation study: CIEE

Scenario	MAF	CIEE		BS		MR		RR		SEM	
		$\widehat{\alpha}_{XY}$	$\widehat{SE}(\widehat{\alpha}_{XY})$	$\widehat{SE}(\widehat{\alpha}_{XY})$		$\widehat{\alpha}_{XY}$	$\widehat{SE}(\widehat{\alpha}_{XY})$	$\widehat{\alpha}_{XY}$	$\widehat{SE}(\widehat{\alpha}_{XY})$	$\widehat{\alpha}_{XY}$	$\widehat{SE}(\widehat{\alpha}_{XY})$
1	0.05	5×10^{-4}	(0.10)	(0.10)	-1×10^{-3}	(0.10)	-1×10^{-3}	(0.10)	3×10^{-4}	(0.10)	
	0.1	7×10^{-4}	(0.07)	(0.07)	-6×10^{-4}	(0.08)	-6×10^{-4}	(0.07)	-2×10^{-4}	(0.07)	
	0.2	1×10^{-4}	(0.06)	(0.06)	-6×10^{-5}	(0.06)	-6×10^{-5}	(0.06)	-6×10^{-4}	(0.06)	
	0.4	3×10^{-4}	(0.05)	(0.05)	-3×10^{-4}	(0.05)	-3×10^{-4}	(0.05)	-4×10^{-5}	(0.05)	
2	0.05	-1×10^{-3}	(0.10)	(0.10)	-2×10^{-3}	(0.10)	-2×10^{-3}	(0.10)	-3×10^{-4}	(0.10)	
	0.1	1×10^{-3}	(0.07)	(0.08)	-1×10^{-4}	(0.08)	-1×10^{-4}	(0.07)	8×10^{-4}	(0.08)	
	0.2	-3×10^{-4}	(0.06)	(0.06)	-2×10^{-4}	(0.06)	-2×10^{-4}	(0.06)	1×10^{-4}	(0.06)	
	0.4	3×10^{-4}	(0.05)	(0.05)	-2×10^{-4}	(0.05)	-2×10^{-4}	(0.05)	8×10^{-4}	(0.05)	
3	0.05	1×10^{-3}	(0.10)	(0.10)	-1×10^{-3}	(0.10)	-1×10^{-3}	(0.10)	1×10^{-3}	(0.10)	
	0.1	1×10^{-3}	(0.07)	(0.07)	-7×10^{-4}	(0.08)	-7×10^{-4}	(0.07)	2×10^{-3}	(0.08)	
	0.2	-1×10^{-4}	(0.06)	(0.06)	-7×10^{-5}	(0.06)	-6×10^{-5}	(0.06)	4×10^{-4}	(0.06)	
	0.4	-1×10^{-4}	(0.05)	(0.05)	-4×10^{-4}	(0.05)	-3×10^{-4}	(0.05)	1×10^{-4}	(0.05)	
4	0.05	-1×10^{-3}	(0.11)	(0.11)	-2×10^{-2}	(0.11)	-2×10^{-2}	(0.11)	-7×10^{-3}	(0.11)	
	0.1	-1×10^{-4}	(0.08)	(0.08)	-2×10^{-2}	(0.08)	-1×10^{-2}	(0.07)	-6×10^{-3}	(0.08)	
	0.2	-2×10^{-3}	(0.06)	(0.06)	-2×10^{-2}	(0.06)	-2×10^{-2}	(0.06)	-6×10^{-3}	(0.06)	
	0.4	-4×10^{-5}	(0.05)	(0.05)	-2×10^{-2}	(0.05)	-2×10^{-2}	(0.05)	-6×10^{-3}	(0.05)	
5	0.05	-2×10^{-3}	(0.11)	(0.11)	-2×10^{-2}	(0.11)	-2×10^{-2}	(0.11)	-6×10^{-3}	(0.11)	
	0.1	-5×10^{-4}	(0.08)	(0.08)	-2×10^{-2}	(0.08)	-2×10^{-2}	(0.07)	-6×10^{-3}	(0.08)	
	0.2	-5×10^{-4}	(0.06)	(0.06)	-2×10^{-2}	(0.06)	-2×10^{-2}	(0.06)	-6×10^{-3}	(0.06)	
	0.4	2×10^{-5}	(0.05)	(0.05)	-2×10^{-2}	(0.05)	-2×10^{-2}	(0.05)	-7×10^{-3}	(0.05)	
6	0.05	1×10^{-3}	(0.11)	(0.11)	-1×10^{-3}	(0.10)	-1×10^{-3}	(0.10)	-2×10^{-2}	(0.11)	
	0.1	7×10^{-4}	(0.08)	(0.08)	-4×10^{-4}	(0.08)	-4×10^{-4}	(0.07)	-2×10^{-2}	(0.08)	
	0.2	-6×10^{-4}	(0.06)	(0.06)	-1×10^{-4}	(0.06)	-1×10^{-4}	(0.06)	-2×10^{-2}	(0.06)	
	0.4	-5×10^{-4}	(0.05)	(0.05)	-6×10^{-4}	(0.05)	-6×10^{-4}	(0.05)	-2×10^{-2}	(0.05)	
7	0.05	4×10^{-4}	(0.13)	(0.13)	-2×10^{-1}	(0.12)	-2×10^{-1}	(0.11)	-1×10^{-1}	(0.12)	
	0.1	-4×10^{-4}	(0.10)	(0.10)	-2×10^{-1}	(0.09)	-2×10^{-1}	(0.08)	-1×10^{-1}	(0.09)	
	0.2	1×10^{-3}	(0.08)	(0.08)	-2×10^{-1}	(0.08)	-1×10^{-1}	(0.06)	-1×10^{-1}	(0.08)	
	0.4	3×10^{-4}	(0.07)	(0.07)	-2×10^{-1}	(0.07)	-1×10^{-1}	(0.05)	-1×10^{-1}	(0.07)	

Supplementary Table A.22: Empirical mean of the direct effect estimates $\widehat{\alpha}_{XY}$ and their standard error estimates obtained through different methods under the null hypothesis in the LM setting, over the $m = 10,000$ replicates. CIEE is the proposed method using estimating equations; BS is CIEE using nonparametric bootstrap standard errors; MR is multiple regression; RR is residual regression; SEM is structural equation modeling. Coefficient estimates obtained from BS are identical to those of CIEE and not shown. Data was generated for $n = 1,000$ individuals. The scenarios when the amount of bias of $\widehat{\alpha}_{XY}$ was larger than 0.01 are highlighted in red.

Scenario	α_{XY}	CIEE		BS	MR		RR		SEM	
		$\overline{\widehat{\alpha}_{XY}}$	$\overline{\widehat{SE}(\widehat{\alpha}_{XY})}$	$\overline{\widehat{SE}(\widehat{\alpha}_{XY})}$	$\overline{\widehat{\alpha}_{XY}}$	$\overline{\widehat{SE}(\widehat{\alpha}_{XY})}$	$\overline{\widehat{\alpha}_{XY}}$	$\overline{\widehat{SE}(\widehat{\alpha}_{XY})}$	$\overline{\widehat{\alpha}_{XY}}$	$\overline{\widehat{SE}(\widehat{\alpha}_{XY})}$
1	0.1	0.10	(0.06)	(0.06)	0.10	(0.06)	0.10	(0.06)	0.10	(0.06)
	0.2	0.20	(0.06)	(0.06)	0.20	(0.06)	0.20	(0.06)	0.20	(0.06)
2	0.1	0.10	(0.06)	(0.06)	0.10	(0.06)	0.10	(0.06)	0.10	(0.06)
	0.2	0.20	(0.06)	(0.06)	0.20	(0.06)	0.20	(0.06)	0.20	(0.06)
3	0.1	0.10	(0.06)	(0.06)	0.10	(0.06)	0.10	(0.06)	0.10	(0.06)
	0.2	0.20	(0.06)	(0.06)	0.20	(0.06)	0.19	(0.06)	0.20	(0.06)
4	0.1	0.10	(0.06)	(0.06)	0.08	(0.06)	0.08	(0.06)	0.09	(0.06)
	0.2	0.20	(0.06)	(0.06)	0.18	(0.06)	0.18	(0.06)	0.19	(0.06)
5	0.1	0.10	(0.06)	(0.06)	0.08	(0.06)	0.08	(0.06)	0.09	(0.06)
	0.2	0.20	(0.06)	(0.06)	0.18	(0.06)	0.18	(0.06)	0.19	(0.06)

Supplementary Table A.23: Empirical mean of the direct effect estimates $\widehat{\alpha}_{XY}$ and their standard error estimates obtained through different methods under the alternative hypotheses in the LM setting, over the $m = 10,000$ replicates. CIEE is the proposed method using estimating equations; BS is CIEE using nonparametric bootstrap standard errors; MR is multiple regression; RR is residual regression. Coefficient estimates from BS are identical to those of CIEE and not shown. Data was generated for $n = 1,000$ individuals for a genetic marker with $\text{MAF}_X = 0.2$. The scenarios where the amount of bias of $\widehat{\alpha}_{XY}$ was larger than or equal to 0.01 are highlighted in red.

Scenario	Censoring	CIEE		BS	G-Estimation	MR	
		$\overline{\widehat{\alpha}_{XY}}$	$\overline{\widehat{SE}(\widehat{\alpha}_{XY})}$	$\overline{\widehat{SE}(\widehat{\alpha}_{XY})}$	$\overline{\widehat{\alpha}_{XY}}$	$\overline{\widehat{\alpha}_{XY}}$	$\overline{\widehat{SE}(\widehat{\alpha}_{XY})}$
1	10%	3×10^{-4}	(0.06)	(0.06)	0.20	4×10^{-4}	(0.06)
	30%	-1×10^{-4}	(0.06)	(0.06)	0.15	-4×10^{-4}	(0.06)
	50%	-3×10^{-5}	(0.07)	(0.07)	0.11	8×10^{-4}	(0.07)
2	10%	-5×10^{-4}	(0.06)	(0.06)	0.25	1×10^{-4}	(0.06)
	30%	-1×10^{-4}	(0.06)	(0.06)	0.20	-5×10^{-4}	(0.06)
	50%	-2×10^{-4}	(0.07)	(0.07)	0.15	-2×10^{-5}	(0.07)
3	10%	-1×10^{-3}	(0.06)	(0.06)	0.25	-2×10^{-4}	(0.06)
	30%	-7×10^{-4}	(0.06)	(0.06)	0.19	1×10^{-3}	(0.06)
	50%	-2×10^{-4}	(0.07)	(0.07)	0.15	1×10^{-3}	(0.07)
4	10%	-3×10^{-4}	(0.06)	(0.06)	0.25	-2×10^{-2}	(0.06)
	30%	3×10^{-4}	(0.06)	(0.06)	0.19	-2×10^{-2}	(0.06)
	50%	-7×10^{-4}	(0.07)	(0.07)	0.14	-2×10^{-2}	(0.07)
5	10%	1×10^{-3}	(0.06)	(0.06)	1×10^{-3}	-2×10^{-2}	(0.06)
	30%	-1×10^{-3}	(0.06)	(0.06)	-2×10^{-3}	-2×10^{-2}	(0.06)
	50%	1×10^{-3}	(0.07)	(0.07)	8×10^{-4}	-2×10^{-2}	(0.07)

Supplementary Table A.24: Empirical mean of the direct effect estimates $\widehat{\alpha}_{XY}$ and their standard error estimates obtained through different methods under the null hypothesis in the AFT setting, over the $m = 10,000$ replicates. CIEE is the proposed method using estimating equations; BS is CIEE using nonparametric bootstrap standard errors; G-Estimation is the sequential G-estimation approach (Lipman et al., 2011); MR is multiple regression. Coefficient estimates from BS are identical to those of CIEE and not shown. Data was generated for $n = 1,000$ individuals for a genetic marker with $\text{MAF}_X = 0.2$ and different censoring rates. The scenarios where the amount of bias of $\widehat{\alpha}_{XY}$ was larger than 0.01 are highlighted in red.

Scenario	α_{XY}	CIEE		BS	MR	
		$\overline{\widehat{\alpha}_{XY}}$	$\overline{\widehat{SE}(\widehat{\alpha}_{XY})}$	$\overline{\widehat{SE}(\widehat{\alpha}_{XY})}$	$\overline{\widehat{\alpha}_{XY}}$	$\overline{\widehat{SE}(\widehat{\alpha}_{XY})}$
1	0.1	0.10	(0.06)	(0.06)	0.10	(0.06)
	0.2	0.20	(0.06)	(0.06)	0.20	(0.06)
2	0.1	0.10	(0.06)	(0.06)	0.10	(0.06)
	0.2	0.20	(0.06)	(0.06)	0.20	(0.06)
3	0.1	0.10	(0.06)	(0.06)	0.10	(0.06)
	0.2	0.20	(0.06)	(0.06)	0.20	(0.06)
4	0.1	0.10	(0.06)	(0.06)	0.08	(0.06)
	0.2	0.20	(0.06)	(0.06)	0.18	(0.06)
5	0.1	0.10	(0.06)	(0.06)	0.08	(0.06)
	0.2	0.20	(0.06)	(0.06)	0.18	(0.06)

Supplementary Table A.25: Empirical mean of the direct effect estimates $\widehat{\alpha}_{XY}$ and their standard error estimates obtained through different methods under the alternative hypotheses in the AFT setting, over the $m = 10,000$ replicates. CIEE is the proposed method using estimating equations; BS is CIEE using nonparametric bootstrap standard errors; MR is multiple regression. Coefficient estimates obtained through BS are identical to those of CIEE and not shown. Data was generated for $n = 1,000$ individuals for a genetic marker with $\text{MAF}_X = 0.2$, with 30% censoring. The scenarios where the amount of bias of $\widehat{\alpha}_{XY}$ was larger than 0.01 are highlighted in red.

Distribution of $Y\|X, K, L, U$	$\overline{\widehat{\alpha}_{XY}}$	$\overline{\widehat{SE}(\widehat{\alpha}_{XY})}$	Type I error
$N(0,1)$	1×10^{-3}	0.08	4.97%
t_8	-3×10^{-4}	0.09	5.17%
t_4	9×10^{-4}	0.11	4.92%
$\log N(0,1)$	6×10^{-4}	0.16	4.87%

Supplementary Table A.26: Results of CIEE under the null hypothesis when the distribution of the primary phenotype is misspecified. Shown are mean estimates of the direct genetic effect, $\overline{\widehat{\alpha}_{XY}}$, mean standard error estimates of the estimated direct genetic effect, $\overline{\widehat{SE}(\widehat{\alpha}_{XY})}$, standard deviation of the direct genetic effect estimates, $SD(\widehat{\alpha}_{XY})$, and empirical type I error estimates obtained through CIEE under the null hypothesis in the LM setting. Data was generated for $n = 1,000$ individuals and $m = 10,000$ replicates for a genetic marker with $\text{MAF}_X = 0.2$, from scenario 7 under the null hypothesis in the LM setting. The distribution of the primary phenotype given covariates in the data generation was chosen to be from a standard normal (cf. Supplementary Table A.22, Table 4.9), t_4, t_8, or log-standard normal distribution.

λ	β_1	β_2	$\overline{\widehat{\beta}_2}$	$SD(\widehat{\beta}_2)$	Proportion of p-values<0.05	
					X=0	X=1
1.0	0.2	0.2	-0.169	0.034	0.553	0.134
	1.0	0.2	-0.150	0.032	0.515	0.088
	0.2	1.0	-0.822	0.121	1	0.755
	1.0	1.0	-0.789	0.125	1	0.444
1.5	0.2	0.2	-0.108	0.024	0.387	0.119
	1.0	0.2	-0.097	0.023	0.362	0.086
	0.2	1.0	-0.465	0.042	1	0.852
	1.0	1.0	-0.439	0.043	1	0.634

Supplementary Table A.27: Bias of parameter estimates and inflated type I errors of the G-estimation approach under the AFT setting. Shown are the empirical mean and empirical standard deviation (SD) of the G-estimation parameter estimates $\widehat{\beta}_2$ under the model $\tilde{T} = \beta_0 + \beta_1 x + \beta_2 k + \varepsilon$ with $E(\varepsilon) = 0$ and $\sqrt{Var(\varepsilon)} > 0$ in section A.4.3. In addition, the proportion of rejected null hypothesis that there is no association between the intermediate phenotype and the adjusted target time-to-event phenotype (i.e., $\beta_2 = 0$) is reported, for each stratum of X (i.e., for $X = 0$ and $X = 1$).

A.5.5 Transcriptomic effects on obesity

	GO ID	Term	Annot	Signif	Expect	Fisher	Classic KS	Elim KS
1	0046903	secretion	1421	64	33.6	3.1×10^{-7}	5.5×10^{-12}	7.1×10^{-2}
2	0045055	regulated exocytosis	714	40	16.88	3.6×10^{-7}	3.4×10^{-12}	2.1×10^{-1}
3	0006887	exocytosis	823	43	19.46	8.4×10^{-7}	1.1×10^{-12}	1.1×10^{-1}
4	0016192	vesicle-mediated transport	1758	73	41.56	9.1×10^{-7}	1.1×10^{-23}	1.6×10^{-3}
5	0002275	myeloid cell activation in immune response	525	32	12.4	9.6×10^{-7}	2.3×10^{-12}	5.7×10^{-2}
6	0008283	cell proliferation	1799	74	42.53	1.1×10^{-6}	3.9×10^{-9}	1.8×10^{-2}
7	1902578	single-organism localization	2956	107	69.89	1.3×10^{-6}	6.2×10^{-12}	3.2×10^{-1}
8	0042127	regulation of cell proliferation	1421	62	33.6	1.5×10^{-6}	2.7×10^{-8}	5.4×10^{-2}
9	0032940	secretion by cell	1308	58	30.92	2.0×10^{-6}	1.4×10^{-12}	1.2×10^{-2}
10	0044699	single-organism process	11716	310	277	3.5×10^{-6}	1.2×10^{-14}	1.3×10^{-1}
11	0002252	immune effector process	1033	48	24.4	5.1×10^{-6}	3.3×10^{-24}	1.9×10^{-2}
12	0009605	response to external stimulus	1842	73	43.6	5.1×10^{-6}	1.1×10^{-9}	1.1×10^{-2}
13	0002274	myeloid leukocyte activation	600	33	14.2	6.0×10^{-6}	7.7×10^{-14}	8.9×10^{-3}
14	0043299	leukocyte degranulation	519	30	12.27	6.2×10^{-6}	3.9×10^{-12}	1.2×10^{-2}
15	0002376	immune system process	2489	91	58.9	7.5×10^{-6}	$< 1.0 \times 10^{-30}$	3.1×10^{-2}
16	0044765	single-organism transport	2819	100	66.65	8.3×10^{-6}	3.3×10^{-11}	5.9×10^{-1}
17	0002283	neutrophil activation in immune response	476	28	11.25	9.2×10^{-6}	7.8×10^{-12}	1
18	0043312	neutrophil degranulation	476	28	11.25	9.2×10^{-6}	7.8×10^{-12}	7.8×10^{-12}
19	0002444	myeloid leukocyte mediated immunity	530	30	12.53	9.3×10^{-6}	6.7×10^{-12}	1.3×10^{-2}
20	0002366	leukocyte activation in immune response	645	34	15.25	1.1×10^{-5}	5.1×10^{-17}	1.0×10^{-2}

Supplementary Table A.28: Results of the GO term enrichment analysis of the 441 genes associated with SAT, top 20 GO terms with smallest p-values based on classic Fisher's test. Shown are the GO term ID and name, the number of annotated genes ('Annot'), number of significant genes ('Signif'), expected number of significant genes ('Expect'), and unadjusted p-values based on the Fisher's exact test as well as classic and elim KS test.

	GO ID	Term	Annot	Signif	Expect	Fisher	Classic KS	Elim KS
1	0002376	immune system process	2489	91	58.85	7.5×10^{-6}	$< 1 \times 10^{-30}$	3.1×10^{-1}
2	0006955	immune response	1716	66	40.57	4.1×10^{-5}	$< 1 \times 10^{-30}$	5.9×10^{-4}
3	0002682	regulation of immune system process	1250	41	29.6	2.1×10^{-2}	3.2×10^{-29}	3.8×10^{-1}
4	0001775	cell activation	1247	51	29.48	7.9×10^{-5}	1.1×10^{-28}	4.1×10^{-1}
5	0045321	leukocyte activation	1104	46	26.1	1.2×10^{-4}	1.7×10^{-28}	1.7×10^{-1}
6	0002252	immune effector process	1033	48	24.4	5.1×10^{-6}	3.3×10^{-24}	1.9×10^{-1}
7	0002684	positive regulation of immune system process	872	29	20.62	4.0×10^{-2}	8.5×10^{-24}	2.4×10^{-1}
8	0016192	vesicle-mediated transport	1758	73	41.56	9.1×10^{-7}	1.1×10^{-23}	1.6×10^{-3}
9	0050776	regulation of immune response	821	30	19.4	1.2×10^{-2}	2.7×10^{-23}	1.5×10^{-4}
10	0051234	establishment of localization	4717	142	112	3.3×10^{-4}	2.0×10^{-20}	5.2×10^{-1}
11	0046649	lymphocyte activation	584	14	13.81	5.2×10^{-1}	8.9×10^{-20}	1.4×10^{-1}
12	0051179	localization	5684	167	134.38	2.2×10^{-4}	9.5×10^{-19}	2.7×10^{-1}
13	0050778	positive regulation of immune response	618	22	14.61	3.7×10^{-2}	1.3×10^{-18}	5.0×10^{-3}
14	0006810	transport	4595	135	108.64	1.5×10^{-3}	2.9×10^{-18}	8.2×10^{-3}
15	0002263	cell activation involved in immune response	647	34	15.3	1.1×10^{-5}	2.6×10^{-17}	5.8×10^{-3}
16	0042110	T cell activation	419	11	9.91	4.0×10^{-1}	3.6×10^{-17}	1.4×10^{-3}
17	0002366	leukocyte activation involved in immune response	645	34	15.25	1.1×10^{-5}	5.1×10^{-17}	1.0×10^{-2}
18	0019882	antigen processing and presentation	222	11	5.25	1.7×10^{-2}	8.9×10^{-17}	7.8×10^{-3}
19	0048002	antigen processing & presentation of peptide antigen	177	6	4.18	2.4×10^{-1}	3.2×10^{-16}	2.8×10^{-2}
20	0050896	response to stimulus	7354	205	173.87	4.8×10^{-4}	5.5×10^{-16}	3.4×10^{-1}

Supplementary Table A.29: Results of the GO term enrichment analysis of the 441 genes associated with SAT, top 20 GO terms with smallest p-values based on classic KS test. Shown are the GO term ID and name, the number of annotated genes ('Annot'), number of significant genes ('Signif'), expected number of significant genes ('Expect'), and unadjusted p-values based on the Fisher's exact test as well as classic and elim KS test.

	GO ID	Term	Annot	Signif	Expect	Fisher	Classic KS	Elim KS
1	0043312	neutrophil degranulation	476	28	11.25	9.2×10^{-6}	7.8×10^{-12}	7.8×10^{-12}
2	0038095	Fc-epsilon receptor signaling pathway	125	2	2.96	8.0×10^{-1}	1.0×10^{-10}	1.0×10^{-10}
3	0031145	anaphase-promoting complex-dependent catabolic process	81	1	1.92	8.6×10^{-1}	1.4×10^{-9}	1.4×10^{-9}
4	0002479	antigen processing & presentation of exogenous peptide antigen via MHC class I	75	0	1.77	1	4.0×10^{-9}	4.0×10^{-9}
5	0050852	T cell receptor signaling pathway	168	1	3.97	9.8×10^{-1}	9.4×10^{-9}	1.1×10^{-8}
6	0051436	negative regulation of ubiquitin-protein ligase activity in mitotic cell cycle	72	1	1.7	8.2×10^{-1}	4.0×10^{-8}	4.0×10^{-8}
7	0038061	NIK/NF-kappaB signaling	120	5	2.84	1.6×10^{-1}	1.2×10^{-9}	4.2×10^{-8}
8	0019886	antigen processing & presentation of exogenous peptide antigen via MHC class II	88	5	2.08	5.7×10^{-2}	9.1×10^{-8}	9.1×10^{-8}
9	0032729	positive regulation of interferon-gamma production	59	4	1.39	5.1×10^{-2}	1.0×10^{-7}	1.0×10^{-7}
10	0033209	tumor necrosis factor-mediated signaling pathway	163	8	3.85	4.0×10^{-2}	1.1×10^{-7}	1.1×10^{-7}
11	0002223	stimulatory C-type lectin receptor signaling pathway	114	1	2.7	9.4×10^{-1}	1.6×10^{-7}	1.6×10^{-7}
12	0030574	collagen catabolic process	63	5	1.49	1.6×10^{-2}	2.1×10^{-7}	2.1×10^{-7}
13	0006521	regulation of cellular amino acid metabolic process	58	1	1.37	7.5×10^{-1}	3.3×10^{-7}	3.3×10^{-7}
14	0051437	positive regulation of ubiquitin-protein ligase activity in regulation of mitotic cell cycle transition	77	1	1.82	8.4×10^{-1}	8.7×10^{-7}	8.7×10^{-7}
15	0031295	T cell costimulation	73	3	1.73	2.5×10^{-1}	9.6×10^{-7}	9.6×10^{-7}
16	0060071	Wnt signaling pathway, planar cell polarity pathway	110	5	2.6	1.2×10^{-1}	1.3×10^{-6}	1.3×10^{-6}
17	0002250	adaptive immune response	316	14	7.47	1.8×10^{-2}	2.9×10^{-11}	1.8×10^{-6}
18	0051301	cell division	546	16	12.91	2.2×10^{-1}	3.3×10^{-8}	1.9×10^{-6}
19	0050690	regulation of defense response to virus by virus	28	3	0.66	2.8×10^{-2}	3.0×10^{-6}	3.0×10^{-6}
20	0031146	SCF-dependent proteasomal ubiquitin-dependent protein catabolic process	72	0	1.7	1	3.3×10^{-6}	3.3×10^{-6}

Supplementary Table A.30: Results of the GO term enrichment analysis of the 441 genes associated with SAT, top 20 GO terms with smallest p-values based on elim KS test. Shown are the GO term ID and name, the number of annotated genes ('Annot'), number of significant genes ('Signif'), expected number of significant genes ('Expect'), and unadjusted p-values based on the Fisher's exact test as well as classic and elim KS test.

	GO ID	Term	Annot	Signif	Expect	Fisher	Classic KS	Elim KS
1	0016054	organic acid catabolic process	223	22	2.8	6.9×10^{-14}	5.50×10^{-14}	7.0×10^{-2}
2	0046395	carboxylic acid catabolic process	223	22	2.8	6.9×10^{-14}	5.50×10^{-14}	7.0×10^{-2}
3	0043436	oxoacid metabolic process	997	44	12.52	1.2×10^{-13}	1.00×10^{-17}	4.4×10^{-1}
4	0006082	organic acid metabolic process	1013	44	12.72	2.0×10^{-13}	7.00×10^{-18}	2.8×10^{-1}
5	0055114	oxidation-reduction process	946	42	11.88	3.9×10^{-13}	9.10×10^{-19}	1.2×10^{-6}
6	0019752	carboxylic acid metabolic process	890	40	11.17	1.1×10^{-12}	9.90×10^{-15}	6.3×10^{-1}
7	0044281	small molecule metabolic process	1925	61	24.17	2.0×10^{-12}	3.70×10^{-28}	3.7×10^{-3}
8	0044282	small molecule catabolic process	329	23	4.13	2.5×10^{-11}	2.10×10^{-12}	6.6×10^{-1}
9	0044710	single-organism metabolic process	3990	89	50.1	8.1×10^{-10}	3.10×10^{-28}	7.2×10^{-3}
10	0009063	cellular amino acid catabolic process	109	13	1.37	1.0×10^{-9}	8.20×10^{-9}	5.0×10^{-2}
11	0032787	monocarboxylic acid metabolic process	512	26	6.43	1.3×10^{-9}	3.70×10^{-15}	1.8×10^{-2}
12	0072329	monocarboxylic acid catabolic process	111	13	1.39	1.3×10^{-9}	5.50×10^{-9}	7.6×10^{-3}
13	0044712	single-organism catabolic process	890	35	11.17	1.3×10^{-9}	1.60×10^{-13}	2.0×10^{-1}
14	0044283	small molecule biosynthetic process	489	25	6.14	2.4×10^{-9}	4.10×10^{-7}	5.5×10^{-4}
15	1901606	alpha-amino acid catabolic process	94	11	1.18	2.5×10^{-8}	2.50×10^{-7}	1.9×10^{-5}
16	0006631	fatty acid metabolic process	318	18	3.99	1.1×10^{-7}	5.00×10^{-10}	1.5×10^{-1}
17	0009083	branched-chain amino acid catabolic process	20	6	0.25	1.2×10^{-7}	3.00×10^{-8}	3.0×10^{-8}
18	0009062	fatty acid catabolic process	91	10	1.14	2.0×10^{-7}	7.50×10^{-9}	3.3×10^{-2}
19	0006635	fatty acid beta-oxidation	70	9	0.88	2.2×10^{-7}	6.70×10^{-9}	1.7×10^{-5}
20	0009081	branched-chain amino acid metabolic process	23	6	0.29	3.1×10^{-7}	3.60×10^{-8}	3.7×10^{-2}

Supplementary Table A.31: Results of the GO term enrichment analysis of the 225 genes associated with $\frac{\text{SAT}}{\text{TAT}}$, top 20 GO terms with smallest p-values based on classic Fisher's test. Shown are the GO term ID and name, the number of annotated genes ('Annot'), number of significant genes ('Signif'), expected number of significant genes ('Expect'), and unadjusted p-values based on the Fisher's exact test as well as classic and elim KS test.

	GO ID	Term	Annot	Signif	Expect	Fisher	Classic KS	Elim KS
1	0044710	single-organism metabolic process	3990	89	50.1	8.1×10^{-10}	3.1×10^{-28}	7.2×10^{-3}
2	0044281	small molecule metabolic process	1925	61	24.17	2.0×10^{-12}	3.7×10^{-28}	3.7×10^{-3}
3	0044699	single-organism process	11716	168	147.11	4.9×10^{-5}	2.9×10^{-24}	8.3×10^{-3}
4	0044763	single-organism cellular process	9699	148	121.78	2.1×10^{-5}	1.2×10^{-23}	7.6×10^{-3}
5	0002376	immune system process	2489	35	31.25	2.6×10^{-1}	2.8×10^{-23}	1.4×10^{-1}
6	0001775	cell activation	1247	13	15.66	7.9×10^{-1}	1.6×10^{-22}	2.0×10^{-1}
7	0006955	immune response	1716	24	21.55	3.2×10^{-1}	7.9×10^{-22}	4.3×10^{-4}
8	0045321	leukocyte activation	1104	10	13.86	8.9×10^{-1}	2.6×10^{-21}	1.2×10^{-1}
9	0055114	oxidation-reduction process	946	42	11.88	3.9×10^{-13}	9.1×10^{-19}	1.2×10^{-6}
10	0006082	organic acid metabolic process	1013	44	12.72	2.0×10^{-13}	7.0×10^{-18}	2.8×10^{-1}
11	0002684	positive regulation of immune system process	872	17	10.95	4.8×10^{-2}	9.8×10^{-18}	5.4×10^{-2}
12	0043436	oxoacid metabolic process	997	44	12.52	1.2×10^{-13}	1.0×10^{-17}	4.4×10^{-1}
13	0002682	regulation of immune system process	1252	17	15.72	4.0×10^{-1}	7.0×10^{-16}	3.4×10^{-1}
14	0032787	monocarboxylic acid metabolic process	512	26	6.43	1.3×10^{-9}	3.7×10^{-15}	1.8×10^{-2}
15	1902578	single-organism localization	2956	48	37.12	3.0×10^{-2}	5.2×10^{-15}	3.9×10^{-1}
16	0006954	inflammatory response	605	20	7.6	7.6×10^{-5}	5.8×10^{-15}	5.4×10^{-13}
17	0019752	carboxylic acid metabolic process	890	40	11.17	1.1×10^{-12}	9.9×10^{-15}	6.3×10^{-1}
18	0044765	single-organism transport	2819	47	35.4	2.1×10^{-2}	1.2×10^{-14}	7.2×10^{-1}
19	0006091	generation of precursor metabolites and energy	343	16	4.31	6.9×10^{-6}	1.8×10^{-14}	2.5×10^{-2}
20	0006629	lipid metabolic process	1252	31	15.72	1.9×10^{-4}	2.5×10^{-14}	2.4×10^{-2}

Supplementary Table A.32: Results of the GO term enrichment analysis of the 225 genes associated with $\frac{\text{SAT}}{\text{TAT}}$, top 20 GO terms with smallest p-values based on classic KS test. Shown are the GO term ID and name, the number of annotated genes ('Annot'), number of significant genes ('Signif'), expected number of significant genes ('Expect'), and unadjusted p-values based on the Fisher's exact test as well as classic and elim KS test.

	GO ID	Term	Annot	Signif	Expect	Fisher	Classic KS	Elim KS
1	0006954	inflammatory response	605	20	7.6	7.6×10^{-5}	5.8×10^{-15}	5.4×10^{-13}
2	0043312	neutrophil degranulation	476	5	5.98	7.2×10^{-1}	1.1×10^{-8}	1.1×10^{-8}
3	0006099	tricarboxylic acid cycle	30	1	0.38	3.2×10^{-1}	1.5×10^{-8}	1.5×10^{-8}
4	0032981	mitochondrial respiratory chain complex I assembly	57	1	0.72	5.1×10^{-1}	2.4×10^{-8}	2.4×10^{-8}
5	0009083	branched-chain amino acid catabolic process	20	6	0.25	1.2×10^{-7}	3.0×10^{-8}	3.0×10^{-8}
6	0002576	platelet degranulation	122	6	1.53	4.4×10^{-3}	8.0×10^{-8}	8.0×10^{-8}
7	0006120	mitochondrial electron transport	46	2	0.58	1.1×10^{-1}	1.2×10^{-7}	1.2×10^{-7}
8	0046321	positive regulation of fatty acid oxidation	14	2	0.18	1.3×10^{-2}	5.3×10^{-7}	5.3×10^{-7}
9	0046487	glyoxylate metabolic process	26	1	0.33	2.8×10^{-1}	9.7×10^{-7}	9.7×10^{-7}
10	0055114	oxidation-reduction process	946	42	11.88	3.9×10^{-13}	9.1×10^{-19}	1.2×10^{-6}
11	0050853	B cell receptor signaling pathway	43	0	0.54	1	3.1×10^{-7}	7.7×10^{-6}
12	0019432	triglyceride biosynthetic process	41	1	0.51	4.0×10^{-1}	8.5×10^{-6}	8.5×10^{-6}
13	0006635	fatty acid beta-oxidation	70	9	0.88	2.2×10^{-7}	6.7×10^{-9}	1.7×10^{-5}
14	1901606	alpha-amino acid catabolic process	94	11	1.18	2.5×10^{-8}	2.5×10^{-7}	1.9×10^{-5}
15	1901136	carbohydrate derivative catabolic process	168	2	2.11	6.3×10^{-1}	7.3×10^{-5}	2.5×10^{-5}
16	0038096	Fc-gamma receptor signaling pathway involved in phagocytosis	72	1	0.9	6.0×10^{-1}	2.8×10^{-5}	2.8×10^{-5}
17	0002690	positive regulation of leukocyte chemotaxis	84	4	1.05	2.1×10^{-2}	3.1×10^{-8}	3.2×10^{-5}
18	0071222	cellular response to lipopolysaccharide	147	4	1.85	1.1×10^{-1}	3.8×10^{-5}	3.8×10^{-5}
19	0042493	response to drug	381	6	4.78	3.4×10^{-1}	3.9×10^{-5}	3.9×10^{-5}
20	0050900	leukocyte migration	380	17	4.77	6.2×10^{-6}	2.8×10^{-11}	6.1×10^{-5}

Supplementary Table A.33: Results of the GO term enrichment analysis of the 225 genes associated with $\frac{\text{SAT}}{\text{TAT}}$, top 20 GO terms with smallest p-values based on elim KS test. Shown are the GO term ID and name, the number of annotated genes ('Annot'), number of significant genes ('Signif'), expected number of significant genes ('Expect'), and unadjusted p-values based on the Fisher's exact test as well as classic and elim KS test.

A.5.6 Study to predict adipokine plasma levels

Measures	Women	Men
Plasma concentration		
Leptin, ng/mL	44.6 (25.2)	18.0 (13.4)
sOB-R, ng/mL	20.0 (8.5)	22.8 (7.1)
Resistin, ng/mL	3.9 (1.1)	4.2 (1.3)
FABP4, ng/mL	32.5 (11.8)	21.3 (7.7)
IL6, pg/mL	1.7 (0.8)	1.9 (0.8)
Total adiponectin, μg/mL	8.5 (3.9)	5.7 (2.0)
HMW adiponectin, μg/mL	4.9 (2.5)	2.8 (1.5)
MMW adiponectin, μg/mL	1.5 (0.9)	1.2 (0.5)
LMW adiponectin, μg/mL	2.0 (1.1)	1.6 (0.5)
SAT gene expression		
Leptin	1×10^{-3} (4×10^{-4})	7×10^{-4} (3×10^{-4})
sOB-R	3×10^{-5} (1×10^{-5})	4×10^{-5} (1×10^{-5})
Resistin	2×10^{-7} (2×10^{-7})	7×10^{-8} (7×10^{-8})
FABP4	3×10^{-2} (6×10^{-3})	3×10^{-2} (6×10^{-3})
IL6	5×10^{-6} (3×10^{-6})	4×10^{-6} (2×10^{-6})
Adiponectin	3×10^{-3} (1×10^{-3})	2×10^{-3} (7×10^{-4})

Supplementary Table A.34: Gender-stratified characteristics of the study population. Values are median and median absolute deviation. Gene expression is shown relative to the housekeeping gene expression in the unit $2^{-\Delta Ct}$, with higher values representing higher expression levels.

	Leptin	sOB-R	Resistin	FABP4	Total Adiponectin	HMW Adiponectin	MMW Adiponectin	LMW Adiponectin	IL6
Weight	0.59	-0.33	-0.05	0.32	-0.31	-0.20	-0.16	-0.11	0.23
BMI	0.62	-0.38	-0.04	0.42	-0.33	-0.24	-0.12	-0.14	0.26
WC	0.62	-0.39	0.00	0.40	-0.28	-0.22	-0.18	-0.16	0.26
HC	0.59	-0.36	0.00	0.35	-0.23	-0.16	-0.09	-0.14	0.23
WHR	0.33	-0.24	0.01	0.25	-0.20	-0.17	-0.20	-0.11	0.17
VAT	0.59	-0.40	-0.06	0.44	-0.28	-0.28	-0.32	-0.10	0.31
CAT	0.51	-0.30	-0.01	0.43	-0.20	-0.28	-0.24	-0.06	0.33
SAT	0.79	-0.37	-0.04	0.50	-0.23	-0.22	-0.08	-0.10	0.27
TAT	0.80	-0.40	-0.05	0.53	-0.25	-0.25	-0.14	-0.12	0.31

Supplementary Table A.35: Partial Pearson correlation coefficients (adjusted for gender, age, physical activity, occupational training) between anthropometric measures and plasma adipokine concentrations. Plasma adipokine concentrations, VAT, CAT, SAT and TAT were log-transformed for the analysis. Correlations of adipokines with the traditional anthropometric measures with r>|0.14| and correlations of adipokines with the MRI-based measures with $r > |0.16|$ correspond to p-values<0.05. Correlations of $r > |0.25|$ and $r > |0.27|$, respectively, correspond to Bonferroni-adjusted p-values<0.05 adjusted for 90 comparisons. HMW, high molecular weight; MMW, medium molecular weight; LMW, low molecular weight.

	Leptin	sOB-R	Resistin	FABP4	Adiponectin	IL6
Weight	0.20	0.13	0.13	-0.21	-0.21	0.20
BMI	0.26	0.11	0.13	-0.19	-0.21	0.22
WC	0.24	0.11	0.13	-0.22	-0.22	0.20
HC	0.28	0.08	0.08	-0.21	-0.19	0.17
WHR	0.08	0.09	0.11	-0.14	-0.14	0.13
VAT	0.25	0.17	0.20	-0.27	-0.30	0.26
CAT	0.22	0.12	0.10	-0.16	-0.21	0.18
SAT	0.37	0.06	0.15	-0.24	-0.25	0.19
TAT	0.35	0.09	0.17	-0.26	-0.28	0.21

Supplementary Table A.36: Partial Pearson correlation coefficients (adjusted for gender, age, physical activity, occupational training) between anthropometric measures and SAT adipokine gene expressions. Gene expression, VAT, CAT, SAT and TAT were log-transformed for the analysis. Correlations of gene expression with the traditional anthropometric measures with r>|0.13| and correlations of gene expression with the MRI-based measures with $r > |0.15|$ correspond to p-values<0.05. Correlations of $r > |0.23|$ and $r > |0.26|$, respectively, correspond to Bonferroni-adjusted p-values<0.05 adjusted for 60 comparisons.

Biomarker	Leptin	sOB-R	Resistin	FABP4	Total Adiponectin	HMW Adiponectin	MMW Adiponectin	LMW Adiponectin	IL6
r	0.52	-0.03	0.11	0.09	0.22	0.24	0.22	0.06	0.15

Supplementary Table A.37: Partial Pearson correlation coefficients r (adjusted for gender) between adipokine plasma concentrations and SAT adipokine gene expressions. Plasma concentrations and SAT gene expressions were log-transformed for the analysis. Correlations $r > |0.15|$ correspond to p-values$<$0.05, and correlations $r > |0.19|$ correspond to Bonferroni-adjusted p-values$<$0.05 adjusted for the 9 comparisons. HMW, high molecular weight; MMW, medium molecular weight; LMW, low molecular.

Model	Variables	Leptin	sOB-R	Resistin	FABP4	Total Adiponectin	HMW Adiponectin	MMW Adiponectin	LMW Adiponectin	IL6
1	GE	0.45	0	0	0.01	0.07	0.06	0.07	0.01	0.01
2	SAT	0.72	0.13	0.01	0.31	0	0	0	0	0.01
3	VAT	0	0.06	0	0.01	0.11	0.12	0.11	0.02	0.12
4	GE×SAT	0.31	0.10	0.02	0.13	0.01	0	0	0	0
5	GE, SAT	0.76	0.12	0.01	0.36	0.02	0.02	0.04	0	0.02
6	GE, SAT, GE×SAT	0.76	0.12	0.01	0.37	0.02	0.01	0.06	0	0.04
7	GE, SAT, VAT	0.77	0.16	0.01	0.37	0.10	0.11	0.12	0.02	0.12

Supplementary Table A.38: Explained variance (adjusted R^2) of the log-plasma levels (PL) by the predictors SAT, VAT and gene expression. Plasma adipokine concentrations, SAT adipokine gene expression, VAT, SAT, and all other other MRI-based measures were log-transformed for the analysis. GE, gene expression; PL, plasma levels; HMW, high molecular weight; MMW, medium molecular weight; LMW, low molecular weight.

Model	Variables	Leptin	sOB-R	Resistin	FABP4	Total Adiponectin	HMW Adiponectin	MMW Adiponectin	LMW Adiponectin	IL6
7	GE, SAT, VAT	0.77	0.16	0.01	0.37	0.10	0.11	0.12	0.02	0.12
8	Sex	0.28	0.01	0.01	0.16	0.08	0.11	0.06	0.04	0.02
9	Sex, GE, SAT, VAT	0.77	0.16	0.01	0.42	0.11	0.13	0.12	0.02	0.12
10	9 + personal	0.77	0.16	0.06	0.42	0.09	0.14	0.17	0.02	0.20
11	10 + other anthr.	0.81	0.22	0.10	0.51	0.09	0.12	0.12	0	0.22
12	10 + other GE	0.79	0.21	0.06	0.48	0.12	0.17	0.25	0.02	0.21
13	10 + other PL	0.77	0.20	0.14	0.47	0.10	0.18	0.19	0.04	0.24
14	10 + other socio	0.78	0.15	0.03	0.43	0.09	0.20	0.13	0	0.19

Supplementary Table A.39: Explained variance (adjusted R^2) of the log-plasma levels (PL) in regression models with additional variables in the left column as predictors, on top of model 7 in Supplementary Table A.38. Plasma adipokine concentrations, SAT adipokine gene expression, VAT, SAT, and all other other MRI-based measures were log-transformed for the analysis. Model 10 includes the variables from model 9 as well as the personal variables age, occupational training, and physical activity. Models 11-14 include the predictors of model 10 in addition to all other MRI-based body compartment measures, all other GE, all other PL, and further personal variables employment status, partner status, smoking status, socioeconomic status and diabetes status. GE, gene expression; PL, plasma levels; HMW, high molecular weight; MMW, medium molecular weight; LMW, low molecular weight.

	Leptin	sOB-R	Resistin	FABP4	Adiponectin	IL6
Leptin	1					
sOB-R	0.24	1				
Resistin	0.02	0.14	1			
FABP4	0.16	0.30	0.05	1		
Adiponectin	0.37	0.22	-0.20	0.64	1	
IL6	0.34	0.30	0.32	0.09	0.06	1

Supplementary Table A.40: Pearson correlation coefficients between the different gene expressions. Gene expressions are shown relative to the housekeeping gene expression in the unit $2^{-\Delta Ct}$ and were log-transformed for the analysis.

	Leptin	sOB-R	Resistin	FABP4	IL6	Total Adiponectin	HMW Adiponectin	MMW Adiponectin	LMW Adiponectin
Leptin	1								
sOB-R	-0.42	1							
Resistin	-0.04	-0.08	1						
FABP4	0.53	-0.18	0.11	1					
IL6	0.02	-0.01	0.18	0.20	1				
Total Adiponectin	-0.02	0.22	-0.11	-0.06	-0.15	1			
HMW Adiponectin	0.01	0.23	-0.15	-0.06	-0.17	0.13	1		
MMW Adiponectin	0.04	0.24	-0.08	0.15	-0.14	0.57	0.59	1	
LMW Adiponectin	-0.04	0.06	-0.02	-0.05	0.01	0.50	0.32	0.00	1

Supplementary Table A.41: Pearson correlation coefficients between the different plasma adipokine concentrations. Plasma concentrations were log-transformed for the analysis. HMW, high molecular weight; MMW, medium molecular weight; LMW, low molecular weight.

	BMI	WC	WHR	TBV	VAT	CAT	SAT	TAT
BMI	1							
WC	0.81	1						
WHR	0.31	0.74	1					
TBV	0.85	0.88	0.51	1				
VAT	0.53	0.80	0.76	0.69	1			
CAT	0.49	0.74	0.72	0.63	0.83	1		
SAT	0.69	0.36	-0.22	0.50	0.11	0.10	1	
TAT	0.80	0.59	0.05	0.68	0.42	0.36	0.95	1

Supplementary Table A.42: Pearson correlation coefficients between the different anthropometric and MRI measures. TBV, VAT, CAT, SAT and TAT were log-transformed for the analysis. TBV, total body volume.

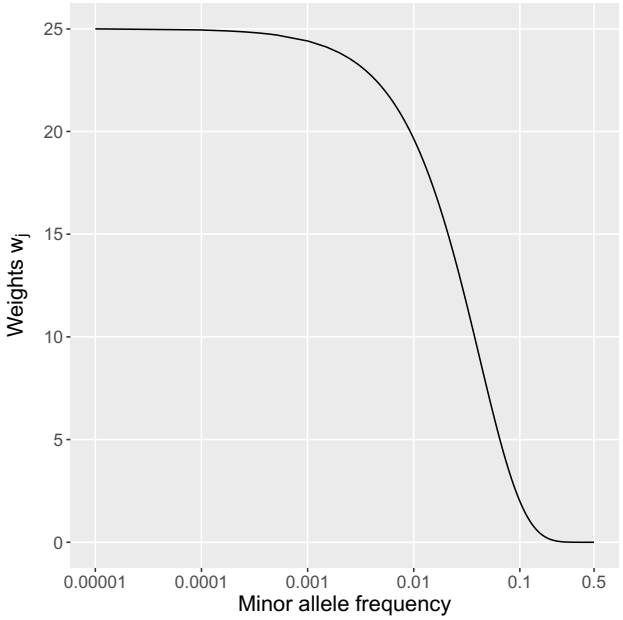
Biomarker	Leptin	sOB-R	Resistin	FABP4	Adiponectin	IL6
Leptin	0.52	0.11	0.21	-0.23	-0.24	0.29
sOB-R	-0.31	-0.03	-0.10	0.21	0.18	-0.21
Resistin	0.08	0.01	0.11	0.01	0.05	0.05
FABP4	0.11	0.14	0.26	0.09	-0.17	0.18
Total Adiponectin	-0.17	-0.13	-0.07	0.19	0.22	-0.07
HMW Adiponectin	-0.17	-0.15	-0.13	0.13	0.20	-0.14
MMW Adiponectin	-0.20	-0.07	-0.03	0.19	0.22	0.03
LMW Adiponectin	-0.11	-0.14	-0.09	0.07	0.06	-0.17
IL6	0.05	0	0.16	0	-0.10	0.15

Supplementary Table A.43: Partial Pearson correlation coefficients (adjusted for gender) between plasma adipokine concentrations and SAT adipokine gene expressions. Plasma concentrations and gene expressions were log-transformed for the analysis. HMW, high molecular weight; MMW, medium molecular weight; LMW, low molecular weight.

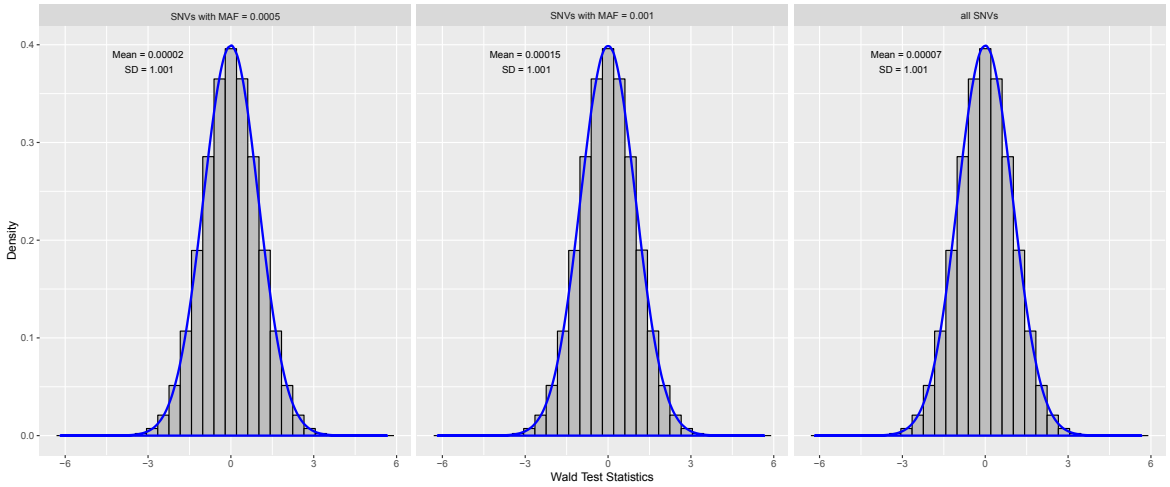
A.6 Supplementary Figures

A.6.1 List of supplementary figures

A.6.2 Simulation study: single-marker versus multi-marker tests

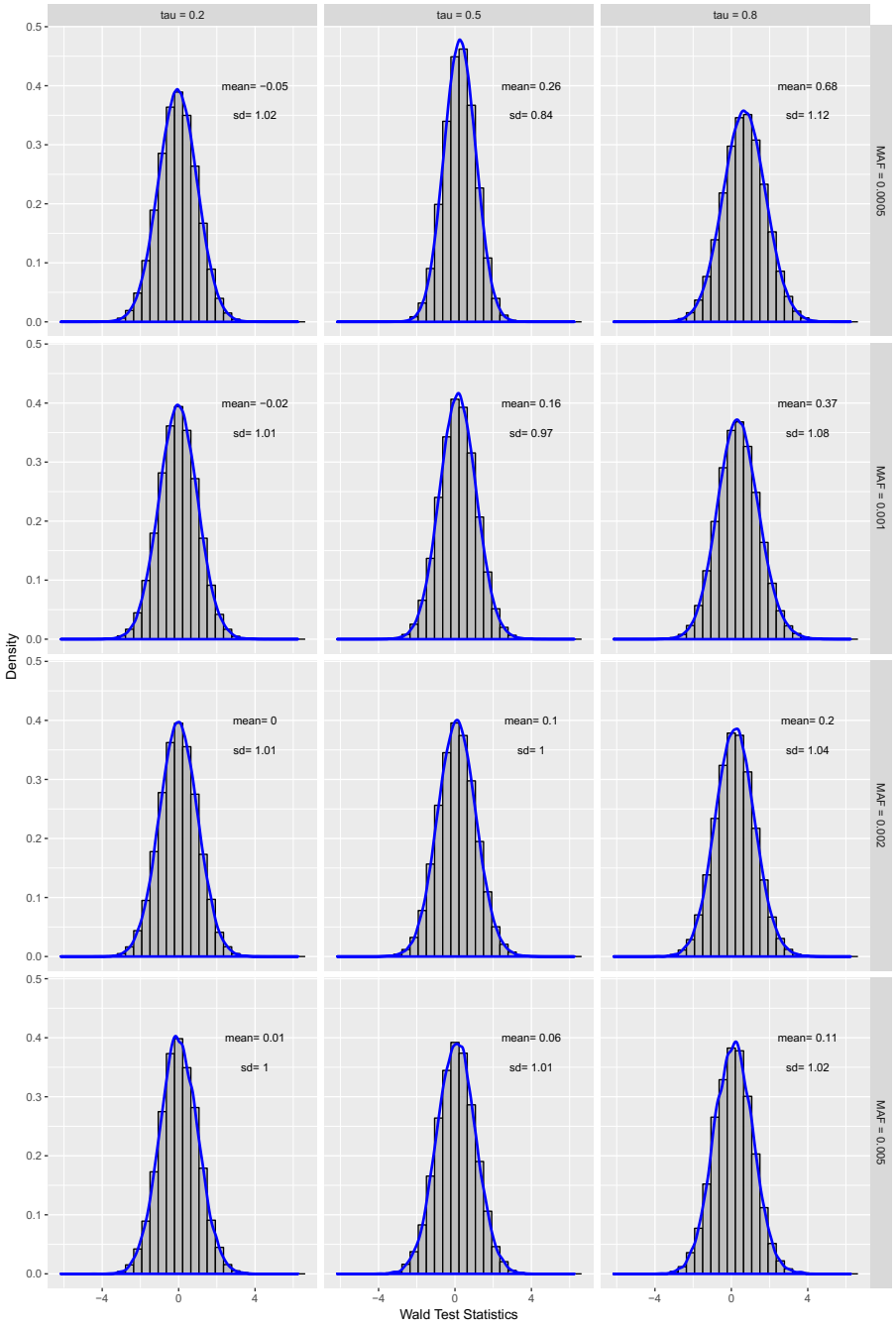


Supplementary Figure A.1: Illustration of the Beta$(1, 25)$ distribution, which is used for the computation of weights w_j of SNVs x_j with a given minor allele frequency in SKAT, SKAT-O, and MURAT.

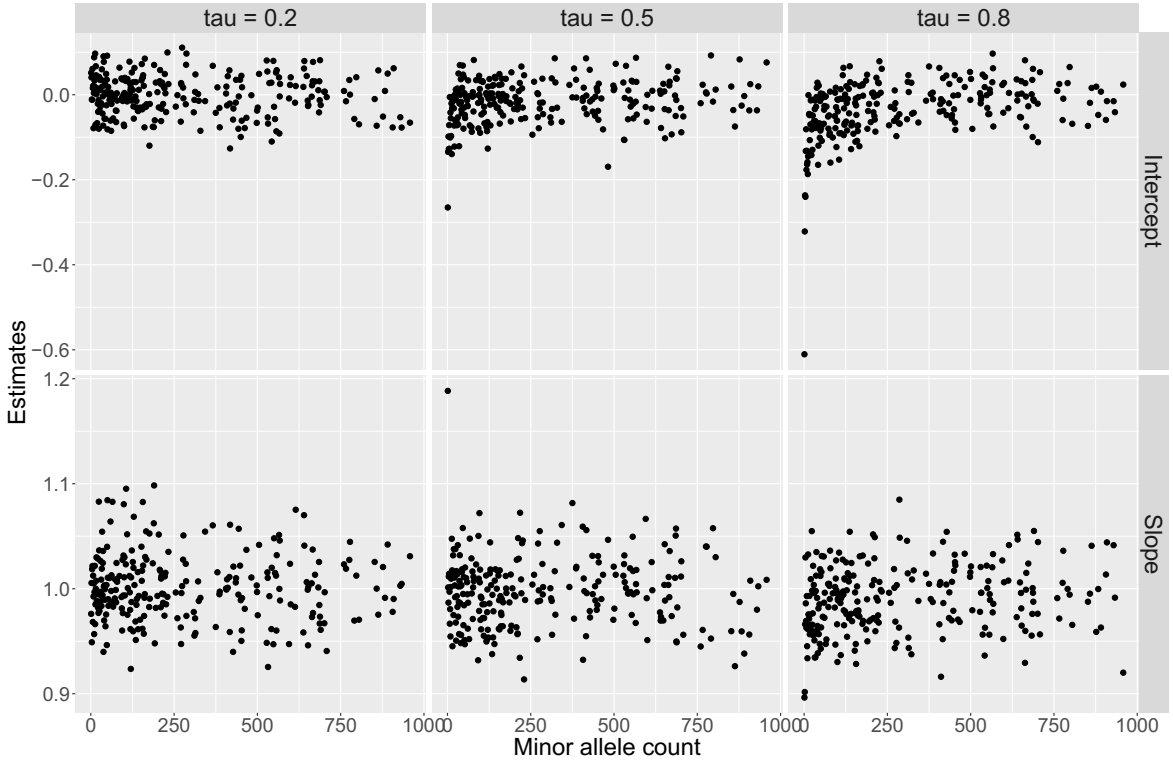


Supplementary Figure A.2: Histogram of the SMT t-test statistic values for singletons, doubletons, and for all SNVs. Datasets were generated from the null model described in scenario 0 in Table 4.1 with size $n = 1,000$ for $m = 10,000,000$ replicates. The histograms and provided descriptive statistics (mean, standard deviation) are based on the 132,797,000 t-test statistics of all singletons in all replicates (left panel), on the 41,341,000 t-test statistics of all doubletons in all replicates (middle panel), and on the 325,393,000 t-test statistics of all SNVs in all replicates (right panel). Overlaid in a blue line is the empirical estimate of the density function. The theoretical mean and standard deviation (SD) of the t_{1000-4} distribution are 0 and $996/994 = 1.001$.

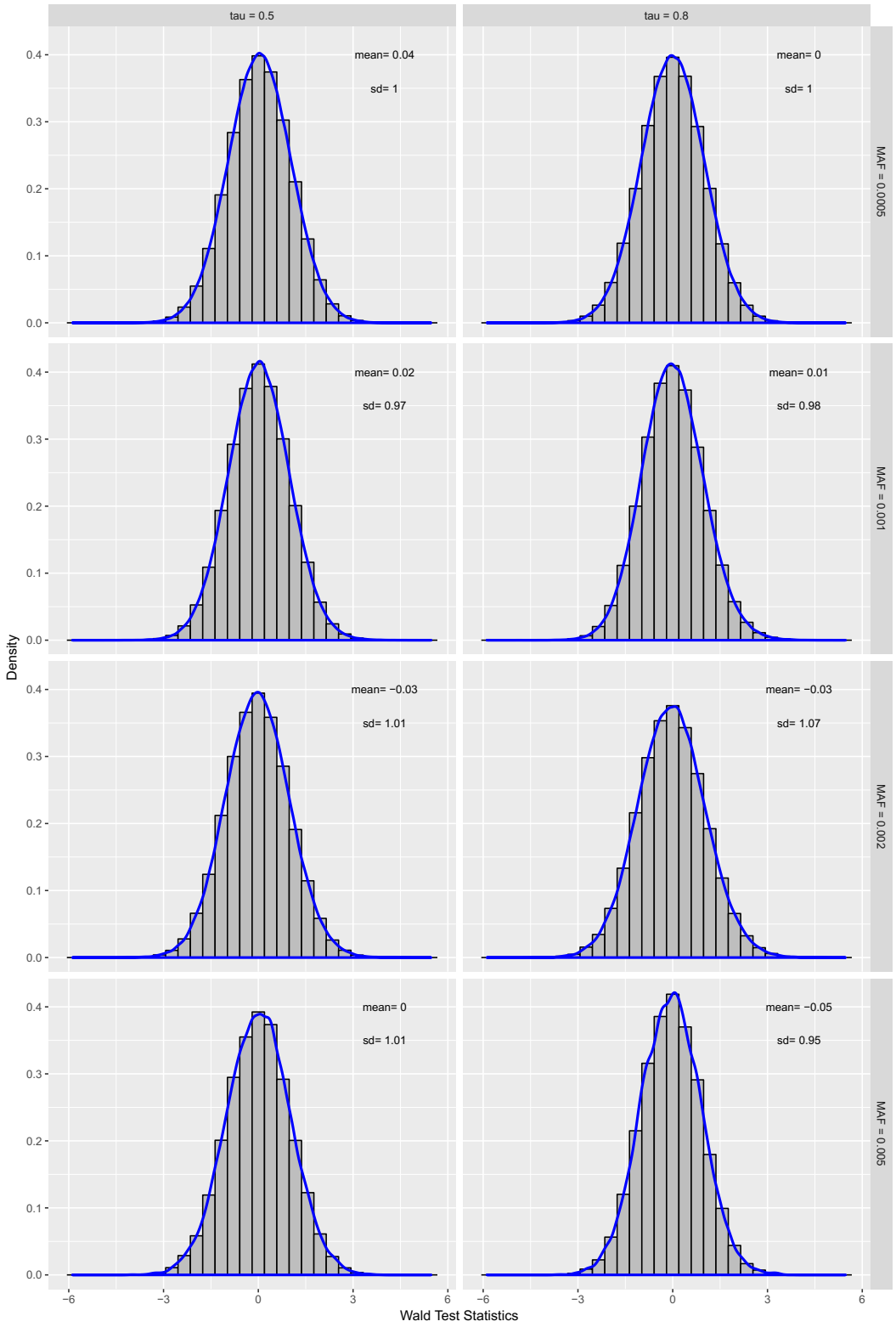
A.6.3 Simulation study: C-JAMP



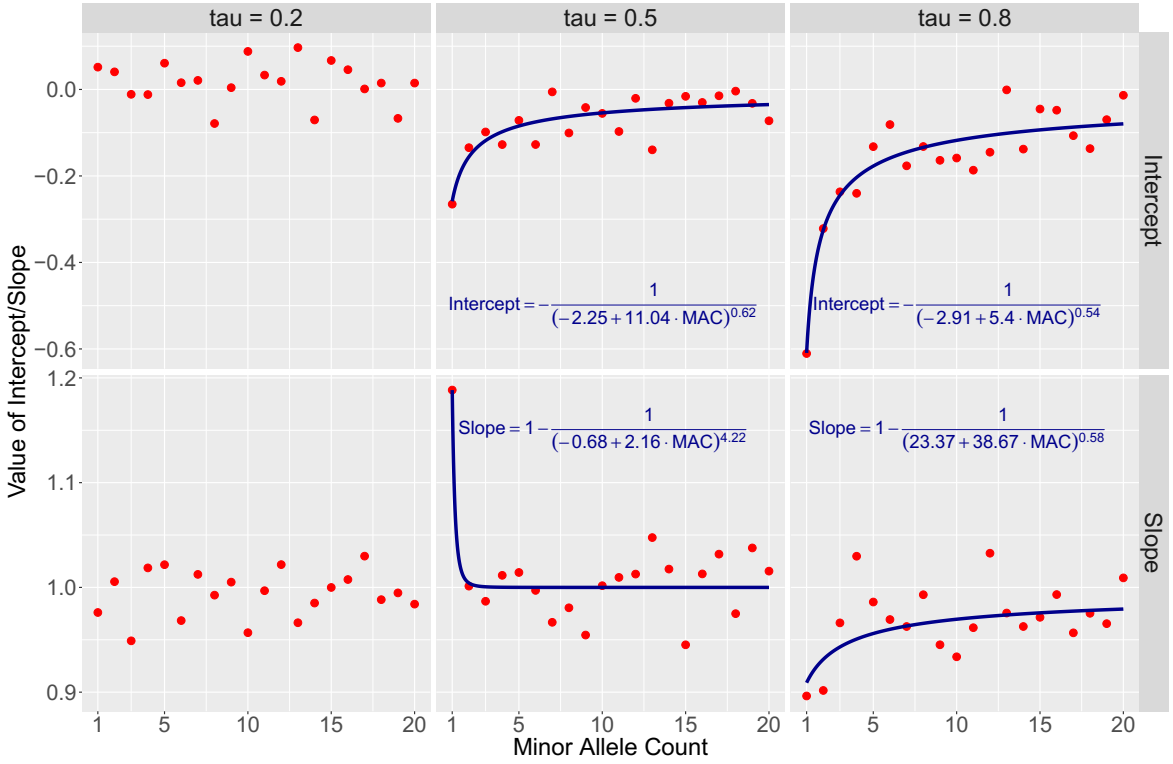
Supplementary Figure A.3: Histograms and density curves of unadjusted C-JAMP Wald test statistics. Data was generated from the null model described in scenario 0 in Table 4.4 for $n = 1,000$ individuals with $m = 100,000$ replicates, and under different dependencies between the two traits Y_1, Y_2 (Kendall's $\tau = 0.2, 0.5, 0.8$). Shown are histograms and density curves of the unadjusted Wald test statistics based on all SNVs with the given MAC in all replicates.



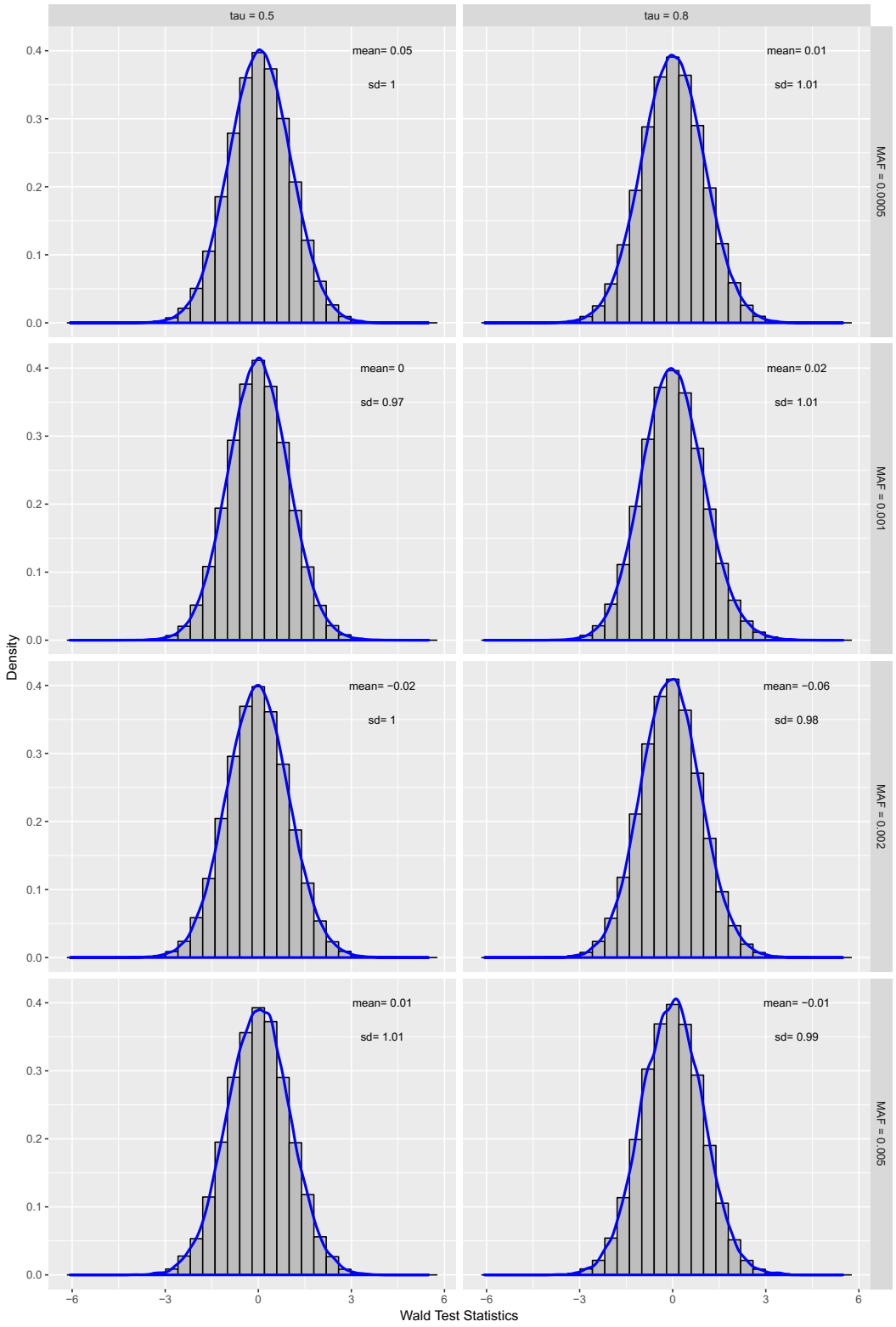
Supplementary Figure A.4: Scatterplots of the intercept/slope parameter estimates to obtain adjusted C-JAMP Wald test statistics by MAC obtained from fitting equation (A.4.2).



Supplementary Figure A.5: Histograms and density curves of adjusted C-JAMP Wald test statistics. Data was generated from the null model described in scenario 0 in Table 4.4 for $n = 1,000$ individuals with $m = 100,000$ replicates, and under different dependencies between the two traits Y_1, Y_2 (Kendall's $\tau = 0.5, 0.8$). Shown are histograms and density curves of the adjusted Wald test statistics based on all SNVs with the given MAC in all replicates.

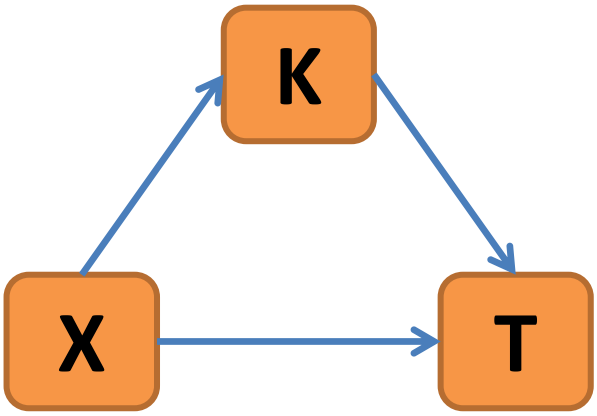


Supplementary Figure A.6: Plot of the functional form of the relationship between MAC and intercept/slope estimates in equation (A.4.2), for different Kendall's τ.



Supplementary Figure A.7: Histograms and density curves of adjusted C-JAMP Wald test statistics using the functional approach through equations (A.4.4)-(A.4.6). Data was generated from the null model described in scenario 0 in Table 4.4 for $n = 1,000$ individuals with $m = 100,000$ replicates, and under different dependencies between the two traits Y_1, Y_2 (Kendall's $\tau = 0.5, 0.8$). Shown are histograms and density curves of the adjusted Wald test statistics based on all SNVs with the given MAC in all replicates.

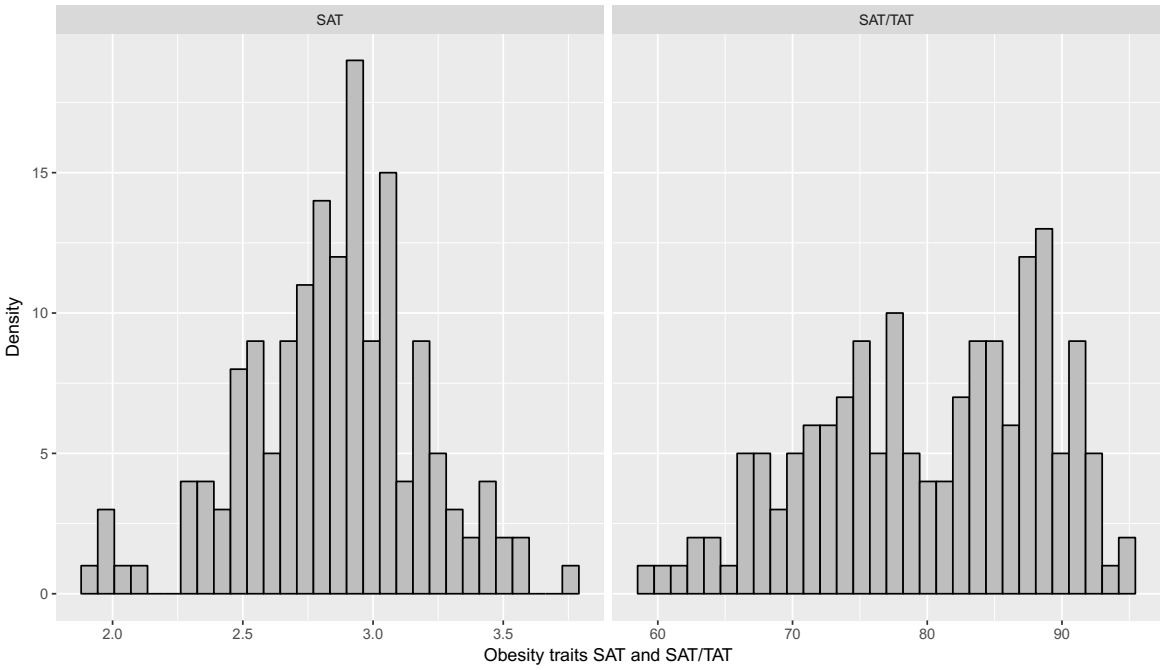
A.6.4 Simulation study: CIEE



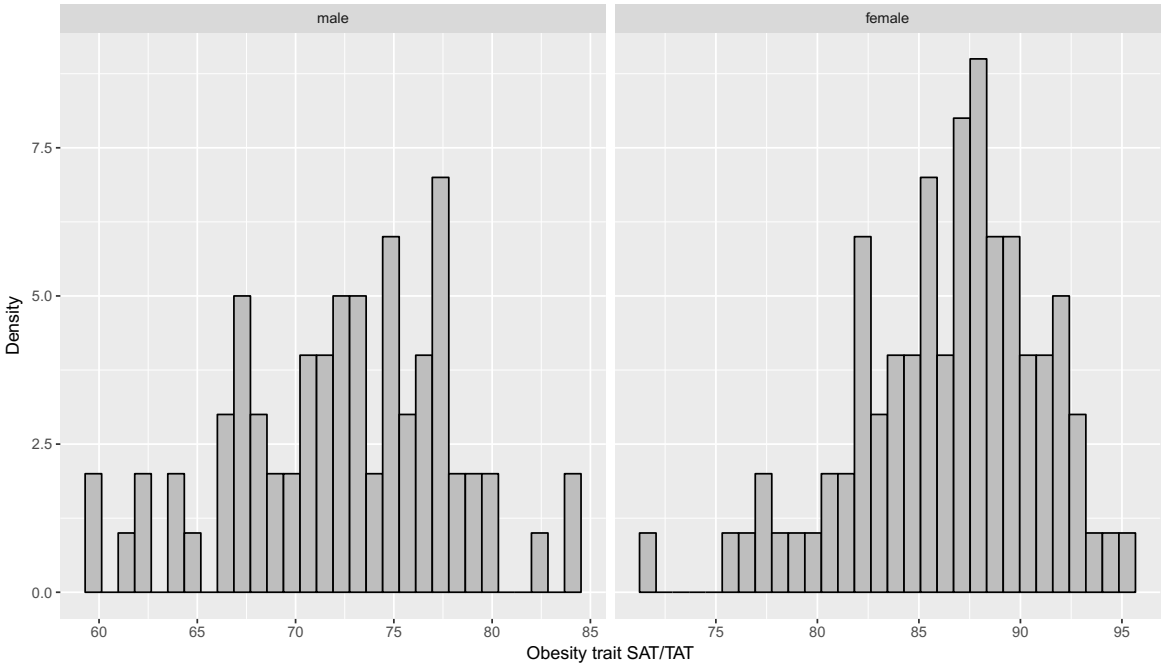
Supplementary Figure A.8: DAG considered in the evaluation of the G-estimation approach described in section A.4.3.

A.6.5 Application: Obesity study

A.6.5.1 Distribution of fat mass and fat distribution measures

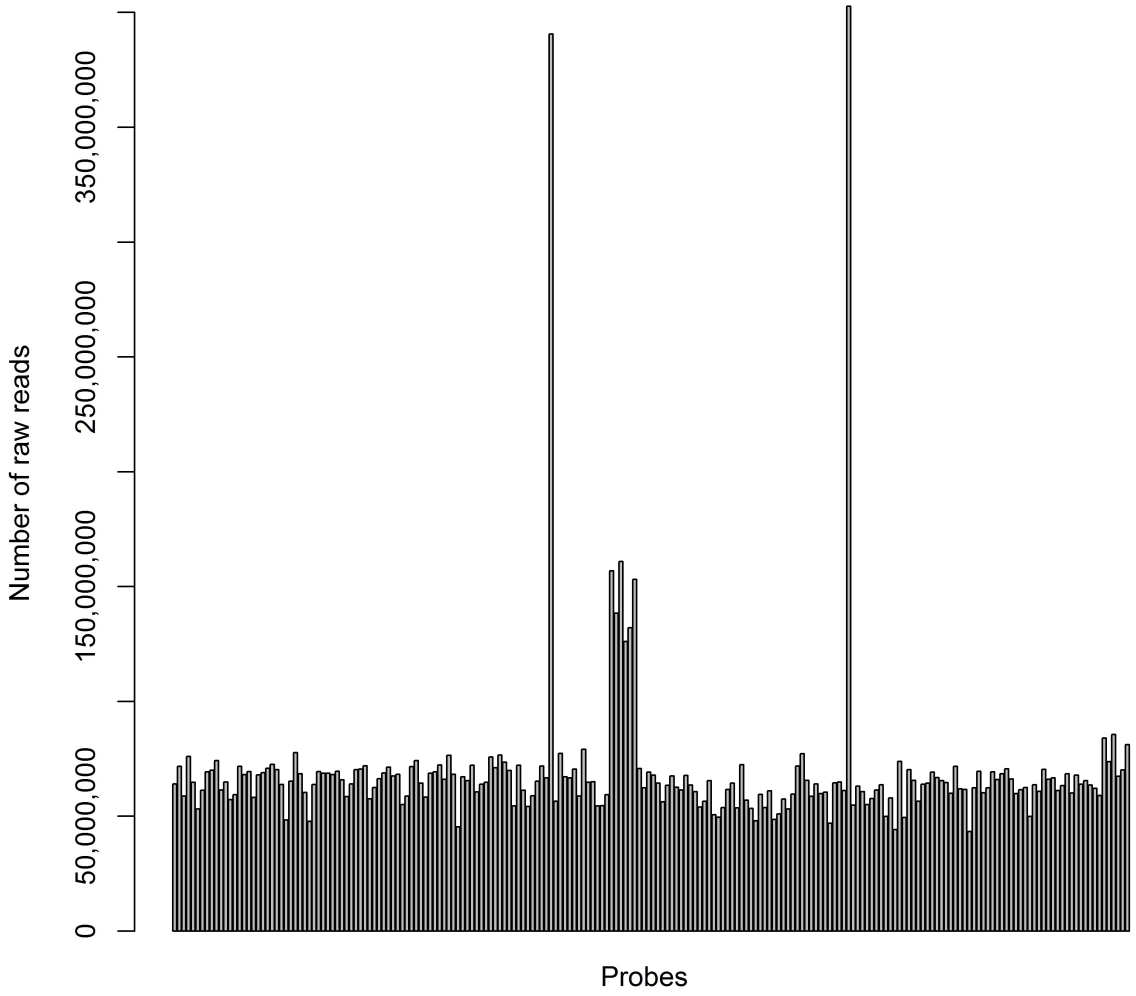


Supplementary Figure A.9: Histograms of $\log(\text{SAT})$ and $\frac{\text{SAT}}{\text{TAT}}$.

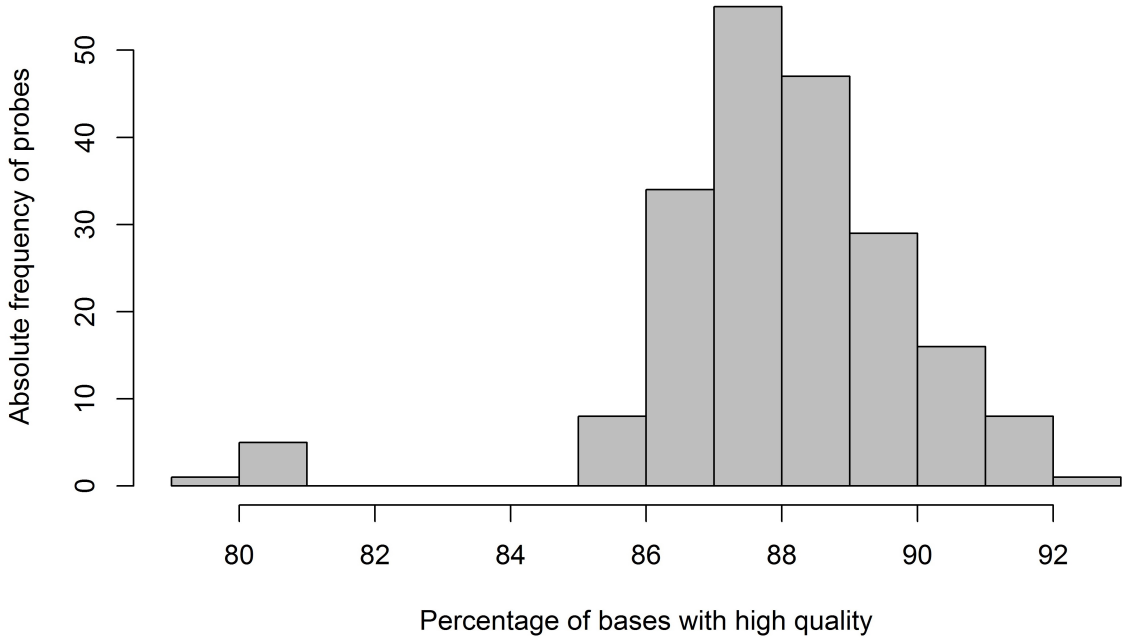


Supplementary Figure A.10: Histograms of $\frac{\text{SAT}}{\text{TAT}}$, stratified by gender.

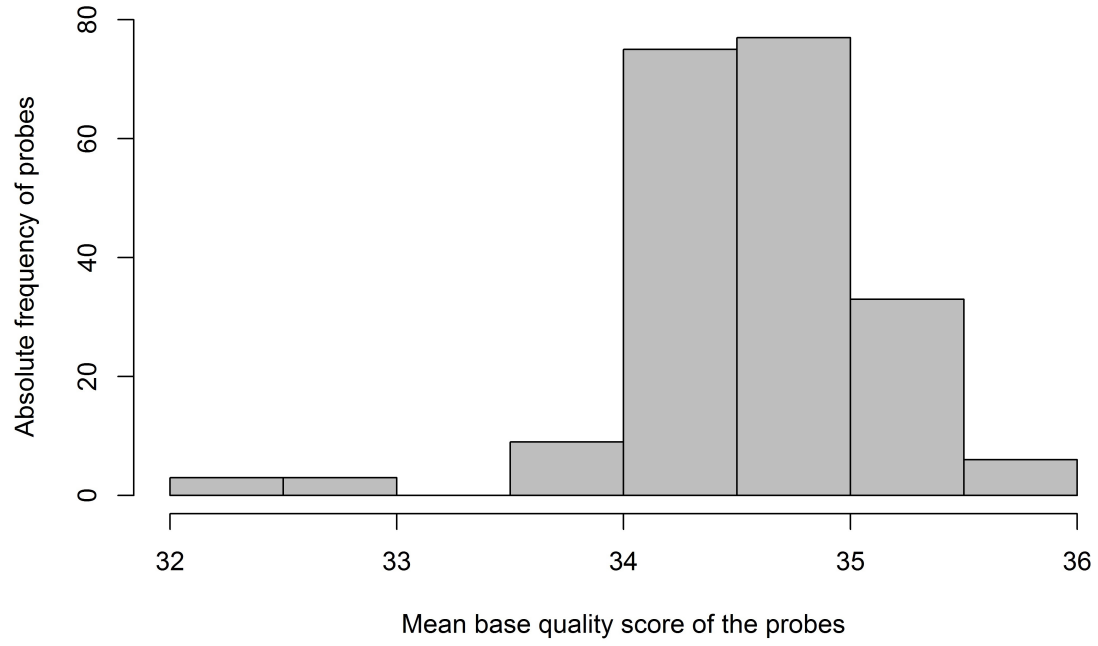
A.6.5.2 Quality control of raw sequencing reads



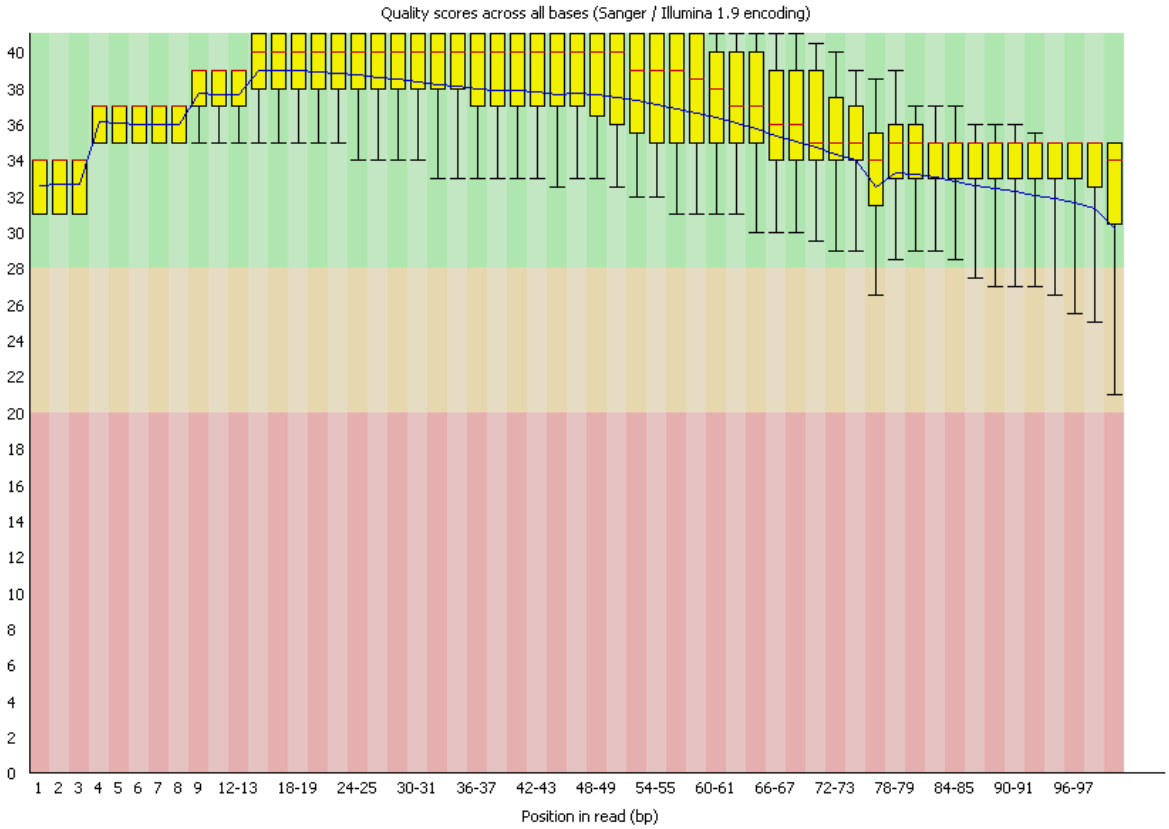
Supplementary Figure A.11: Barplot of the number of raw sequencing reads.



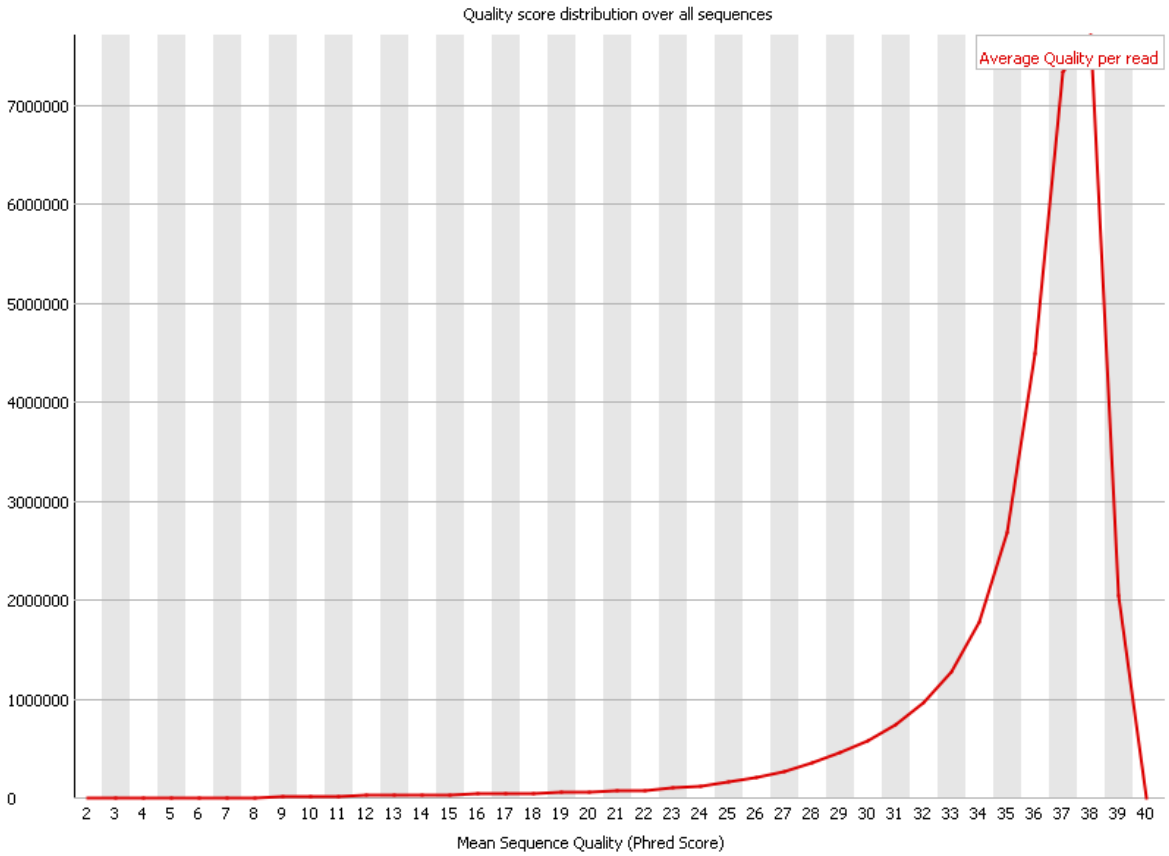
Supplementary Figure A.12: Histogram of the samples' percentage of bases with high quality.



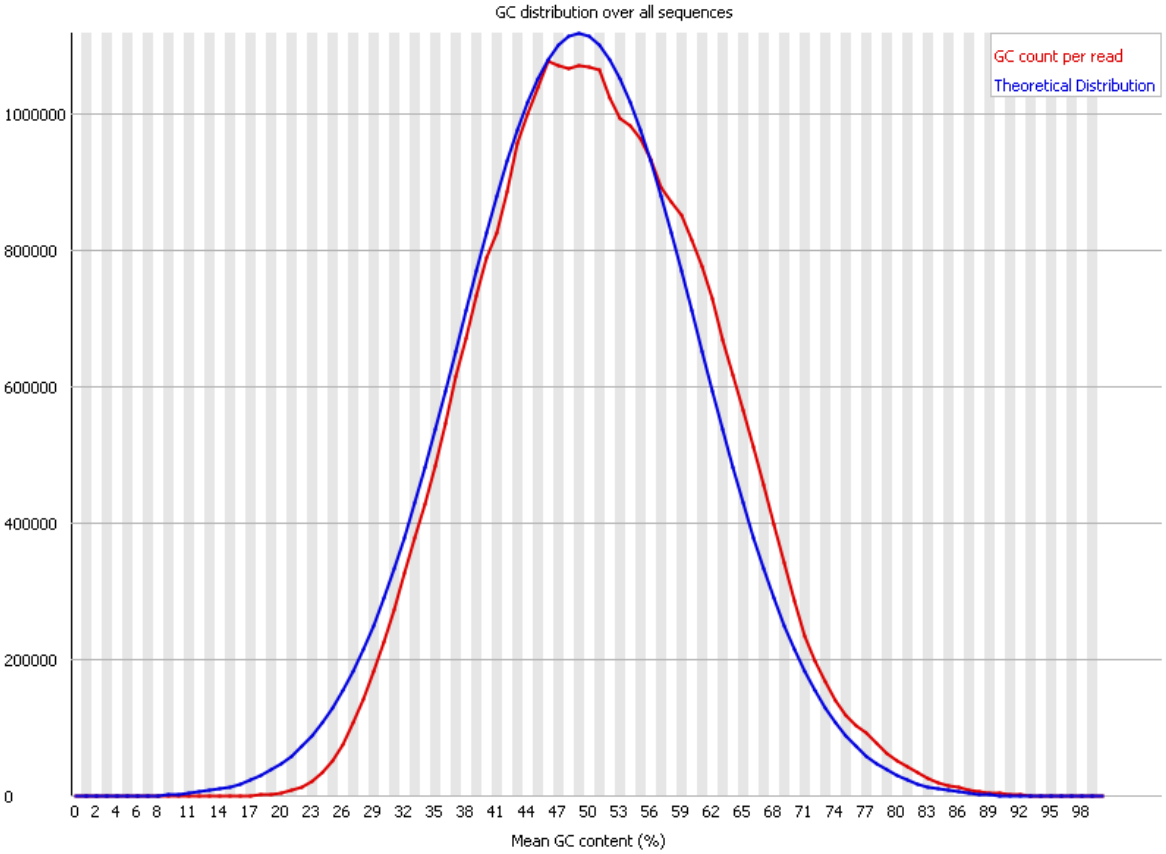
Supplementary Figure A.13: Histogram of the mean base quality score of the reads.



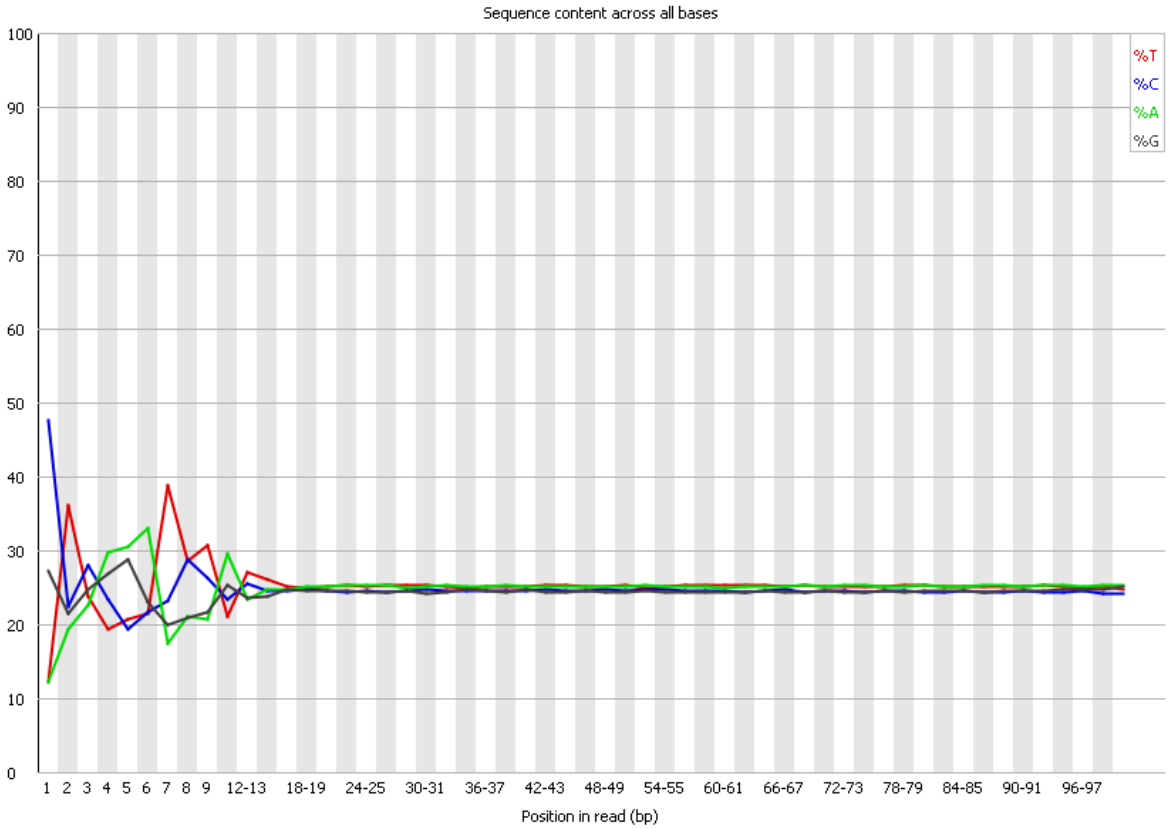
Supplementary Figure A.14: Quality scores across base pairs of all reads, from one exemplary probe.



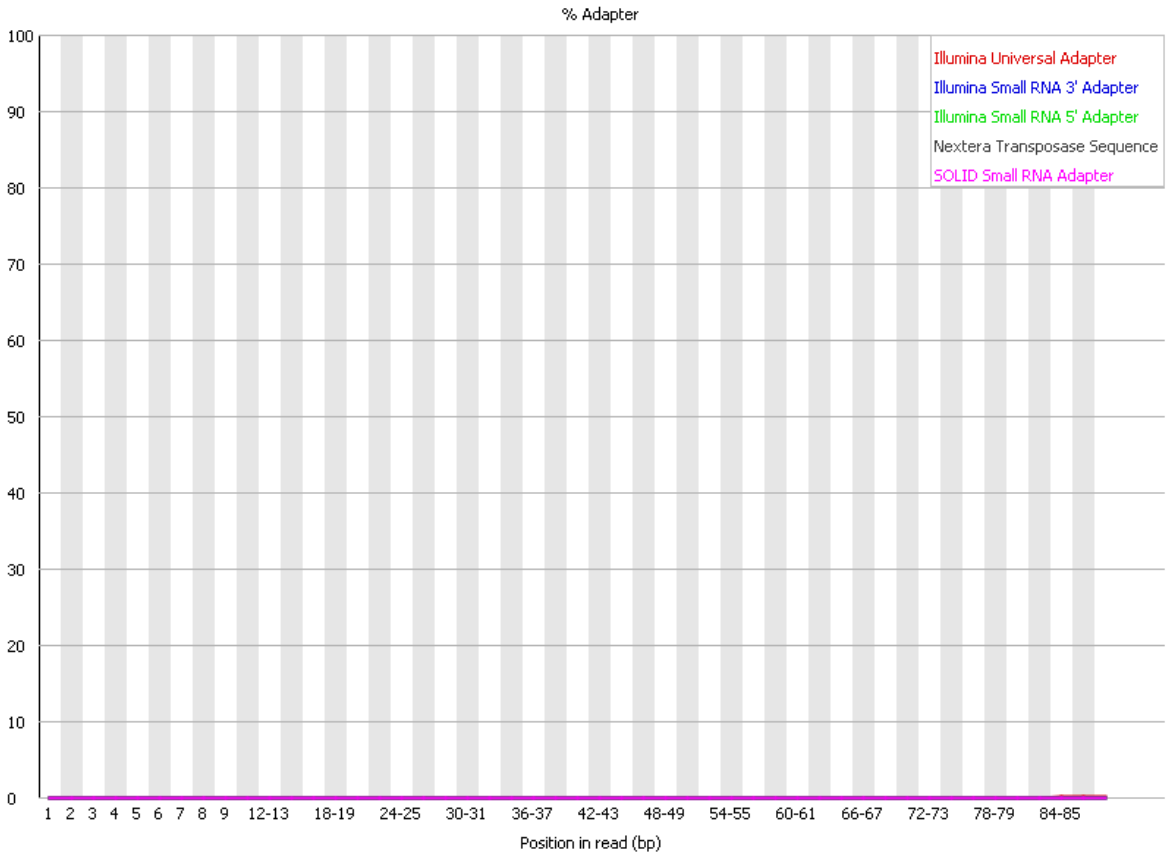
Supplementary Figure A.15: Distribution of quality scores across all reads, from one exemplary probe.



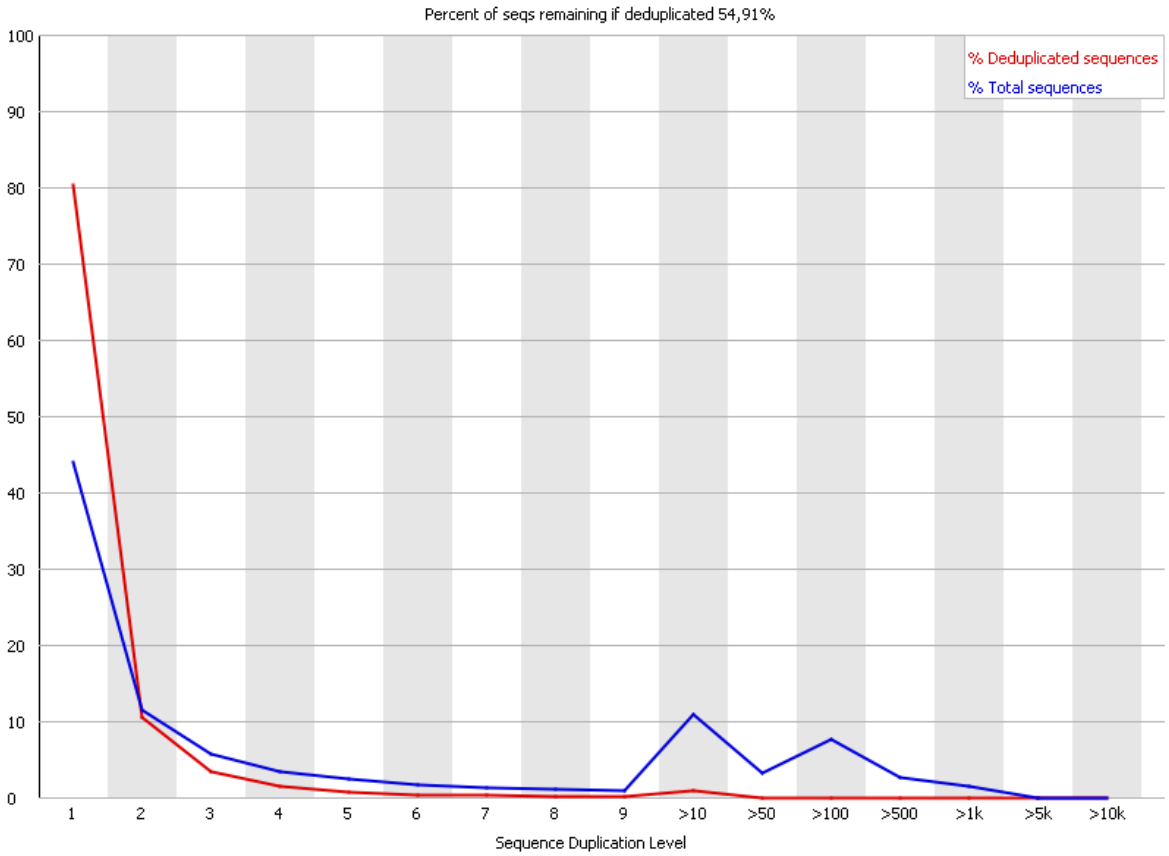
Supplementary Figure A.16: GC distribution across all reads, from one exemplary probe.



Supplementary Figure A.17: Sequence content across all bases, from one exemplary probe.

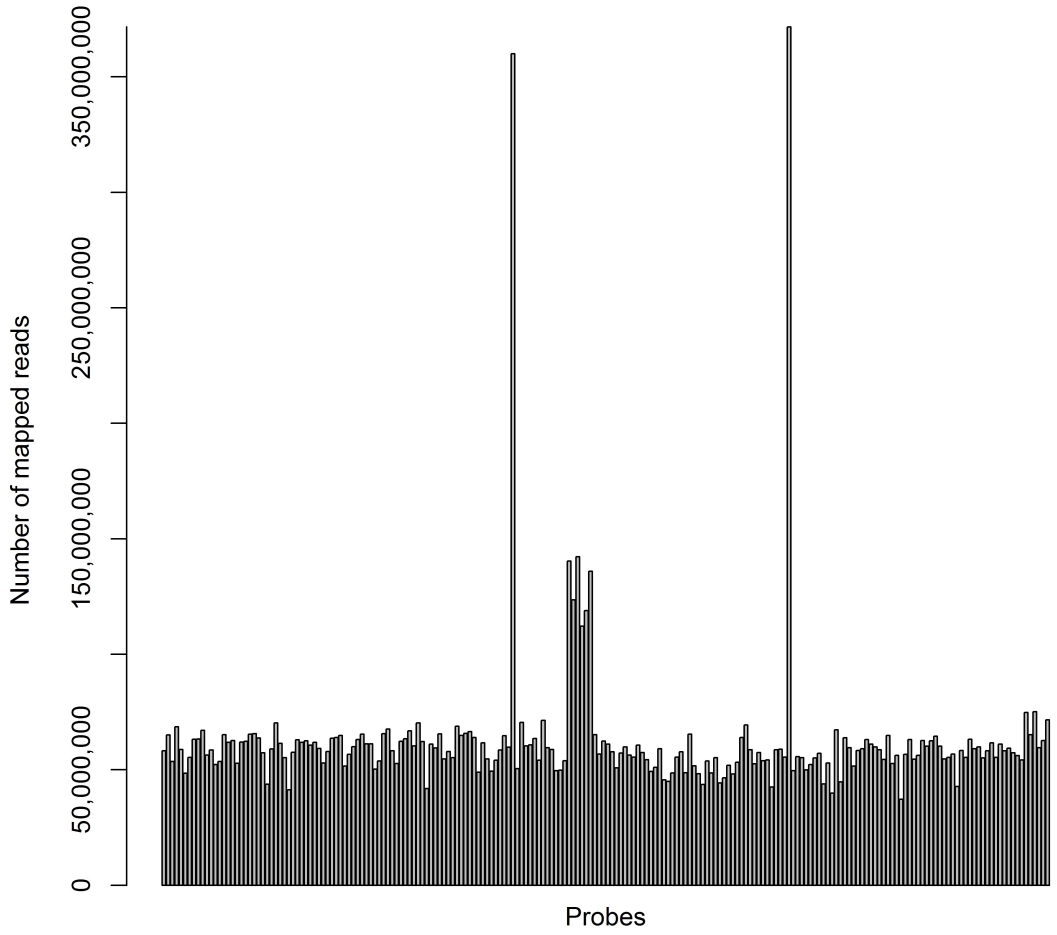


Supplementary Figure A.18: Adapter percentage across all bases, from one exemplary probe.

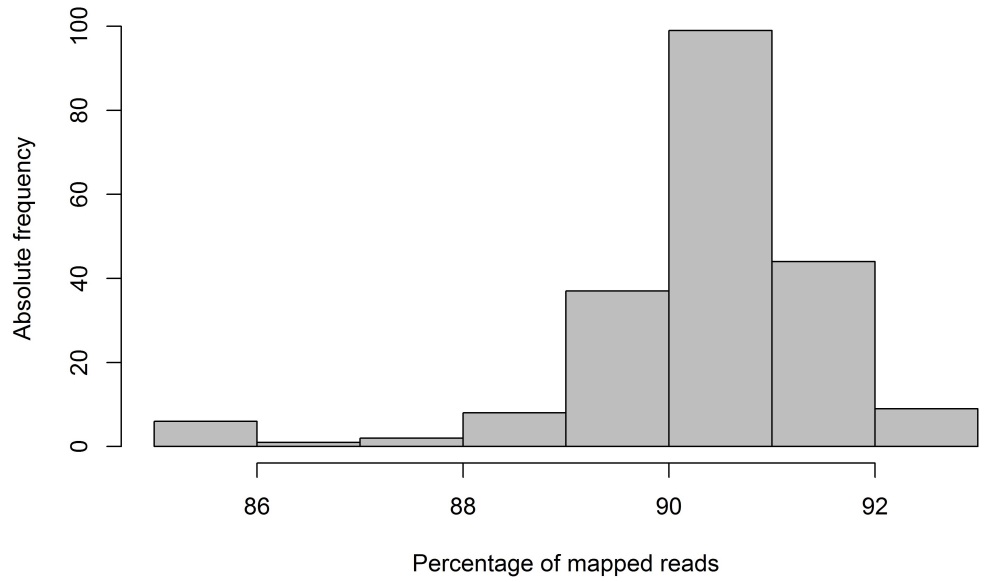


Supplementary Figure A.19: Percentage of duplicated sequences, from one exemplary probe.

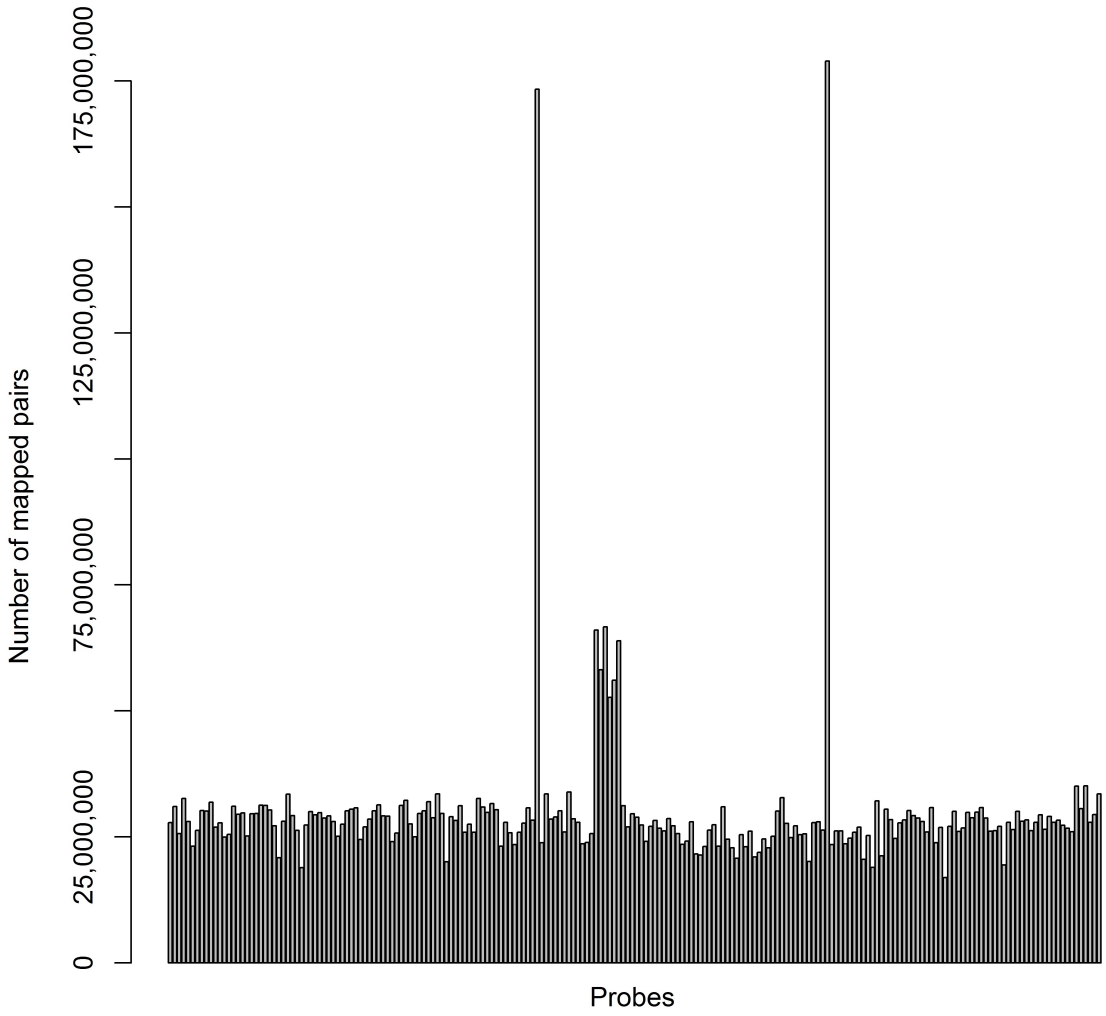
A.6.5.3 Quality control of aligned sequencing reads



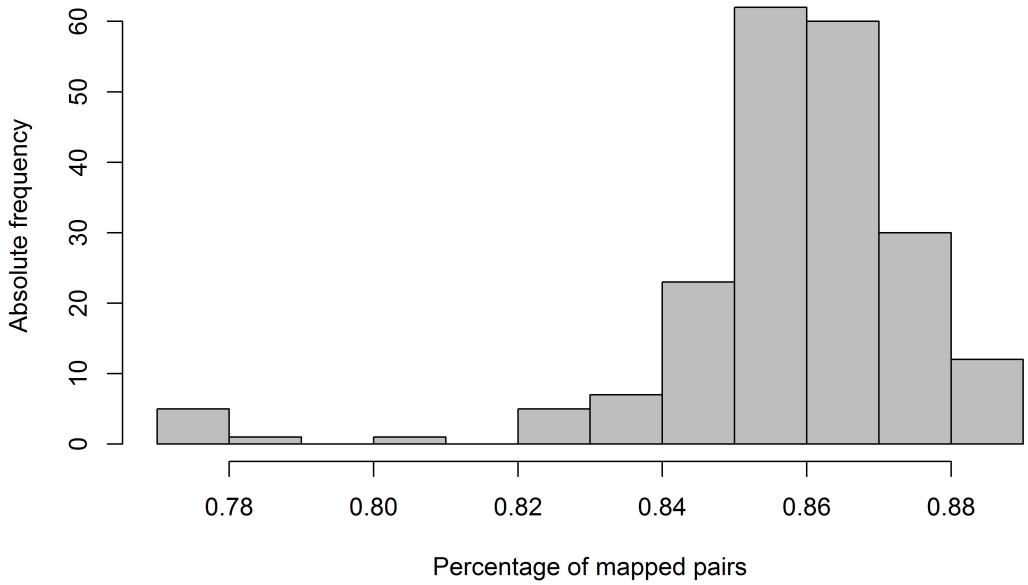
Supplementary Figure A.20: Barplot of the number of mapped reads.



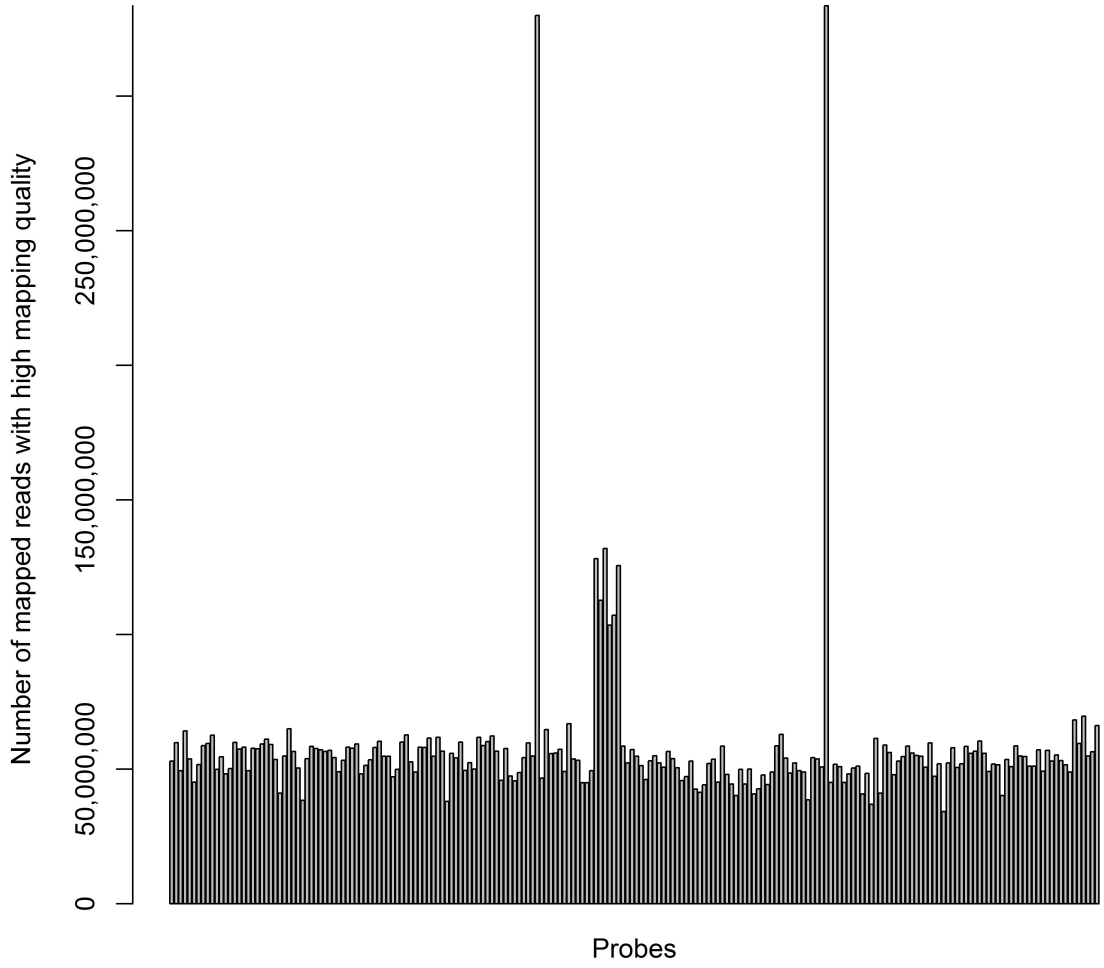
Supplementary Figure A.21: Histogram of the percentage of mapped reads.



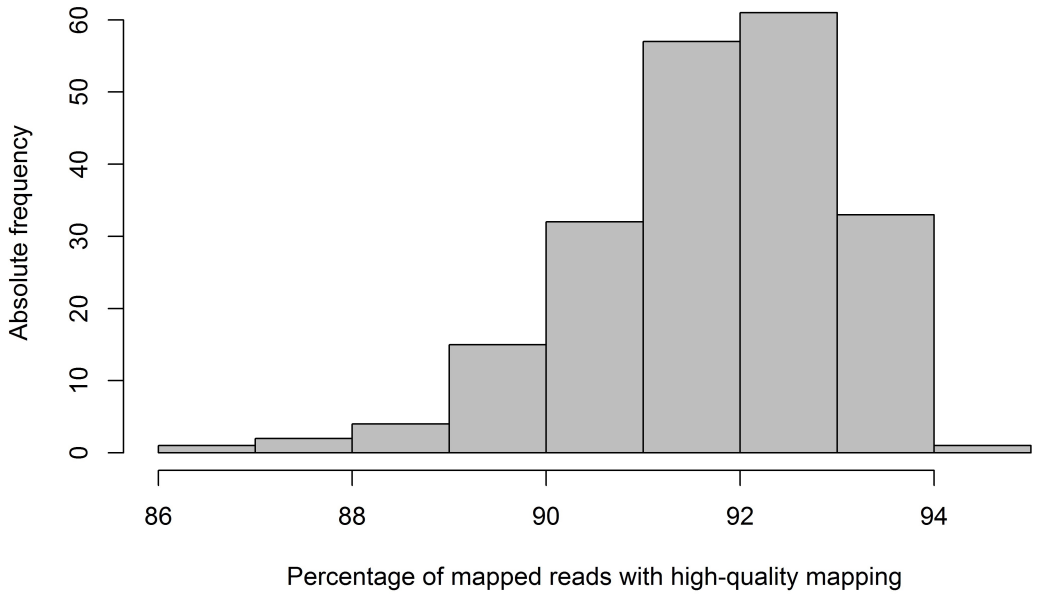
Supplementary Figure A.22: Barplot of the number of mapped pairs.



Supplementary Figure A.23: Histogram of the percentage of mapped pairs.

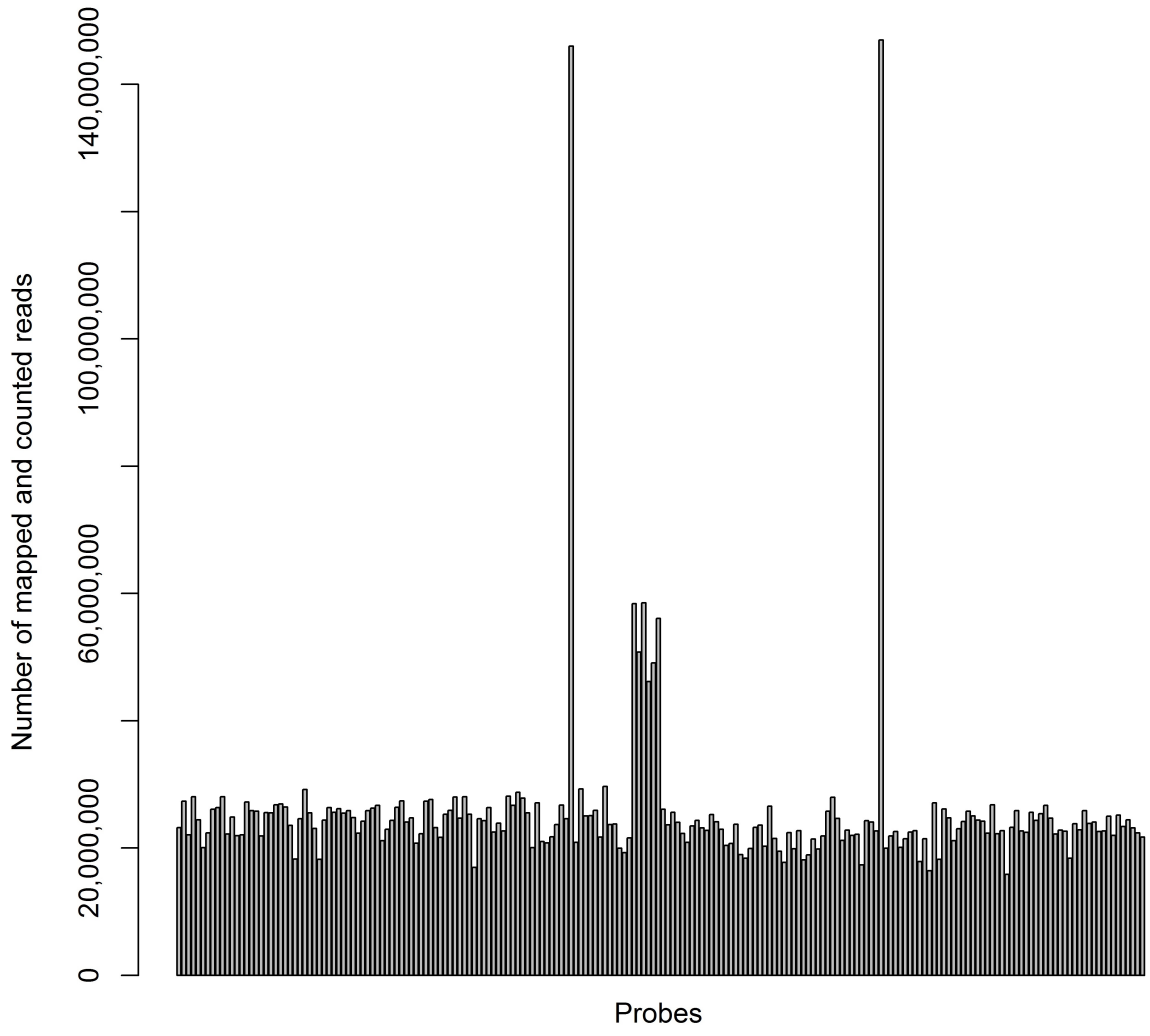


Supplementary Figure A.24: Barplot of the number of mapped reads with high-quality mapping.

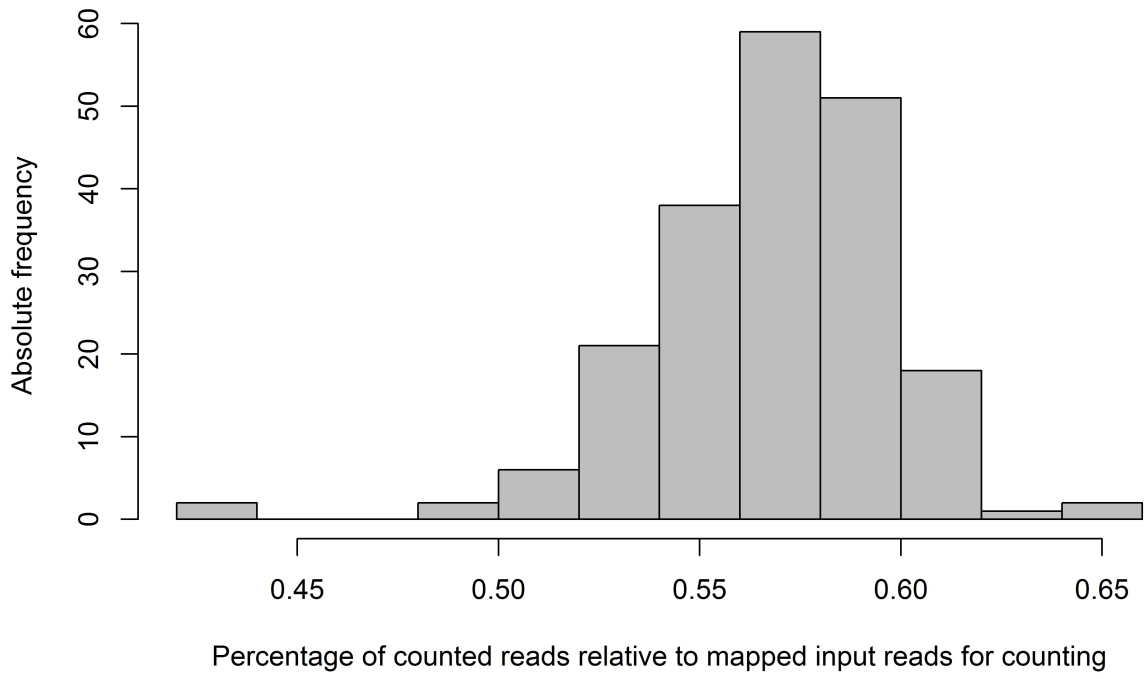


Supplementary Figure A.25: Histogram of the percentage of high-quality mapped reads.

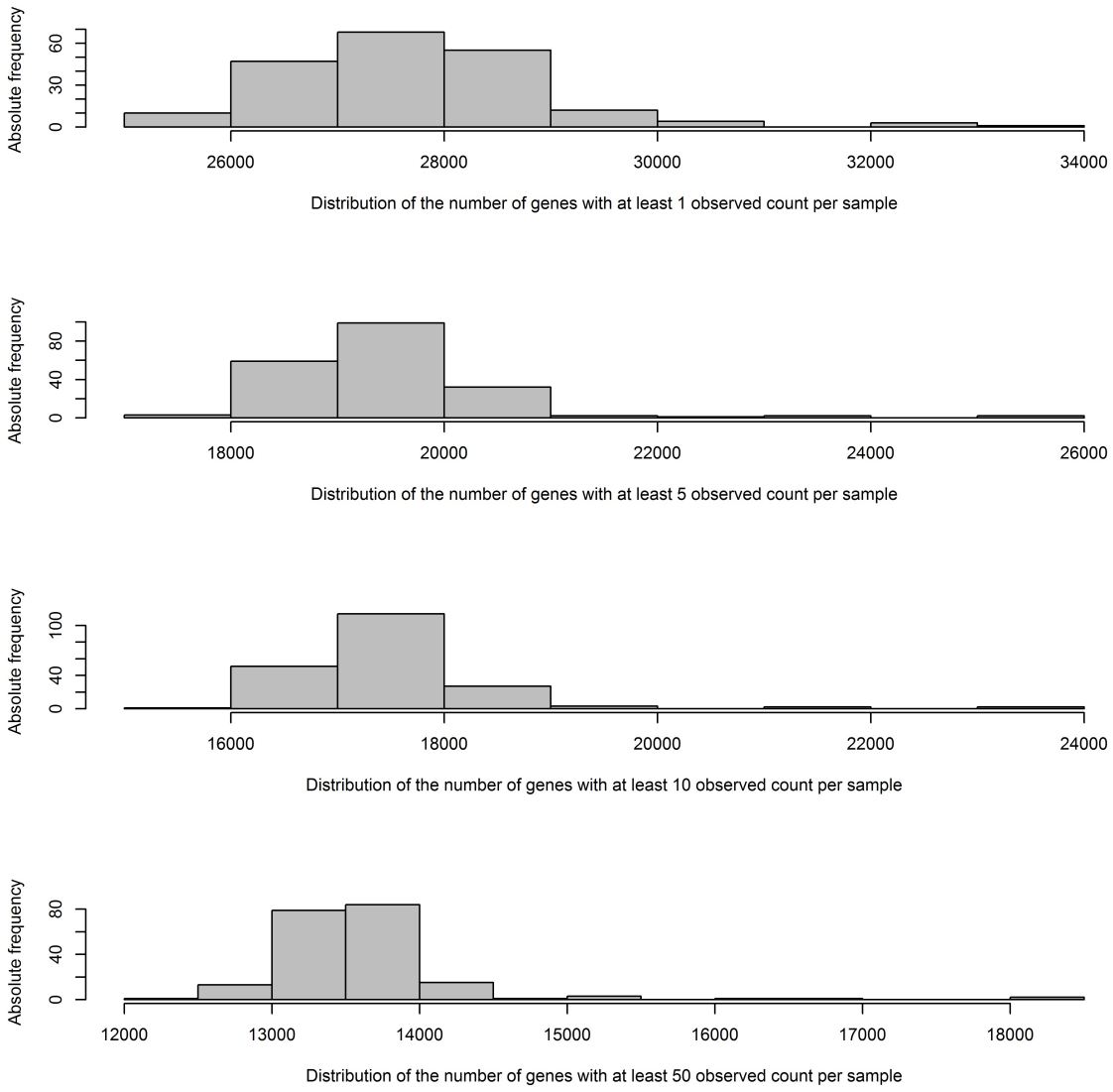
A.6.5.4 Descriptive statistics and quality control of raw and normalized read counts



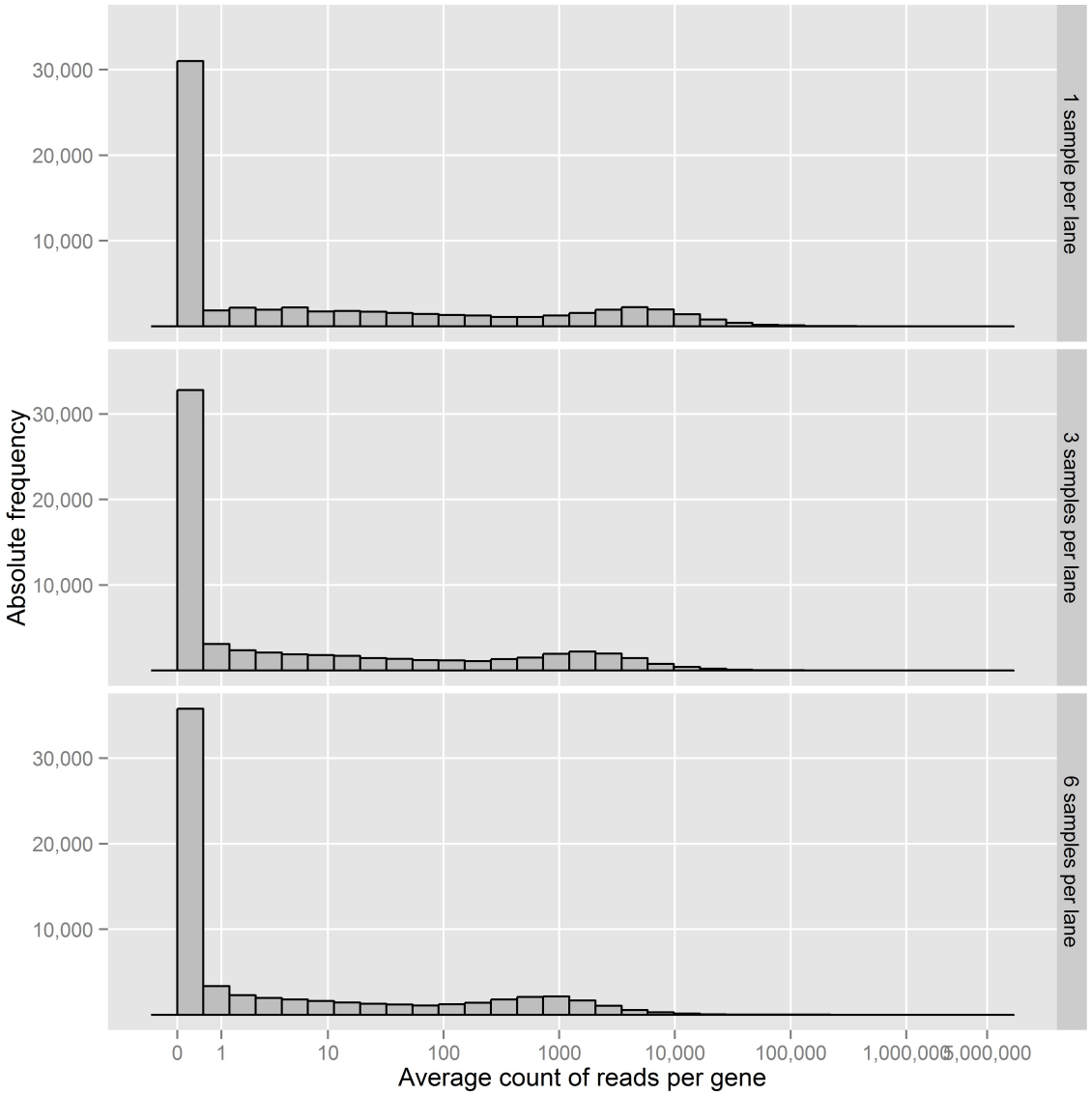
Supplementary Figure A.26: Barplot of the number of mapped and counted reads.



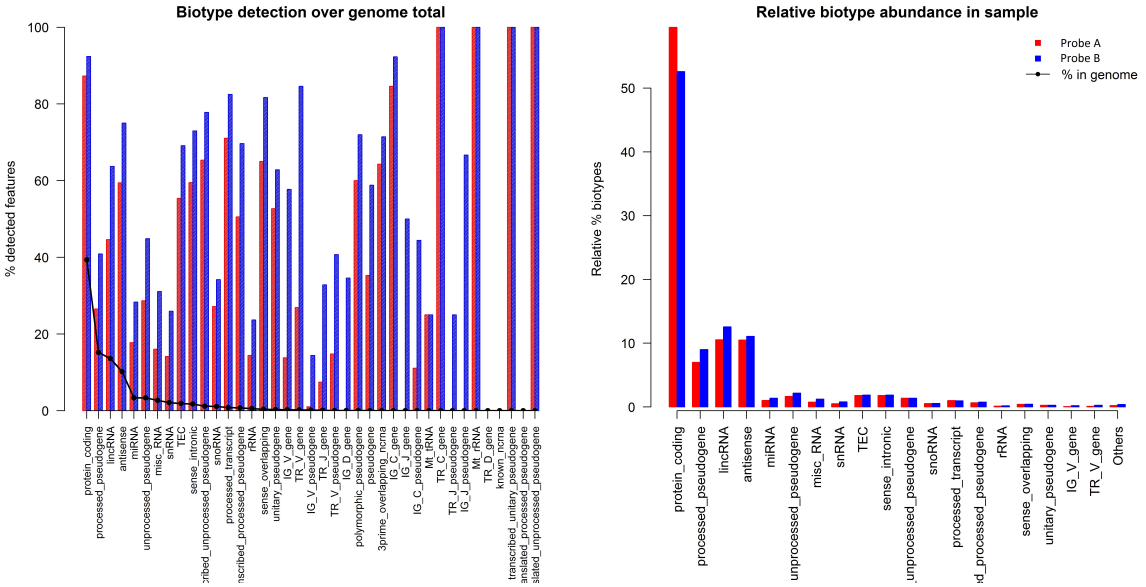
Supplementary Figure A.27: Histogram of the percentage of the number of mapped and counted reads.



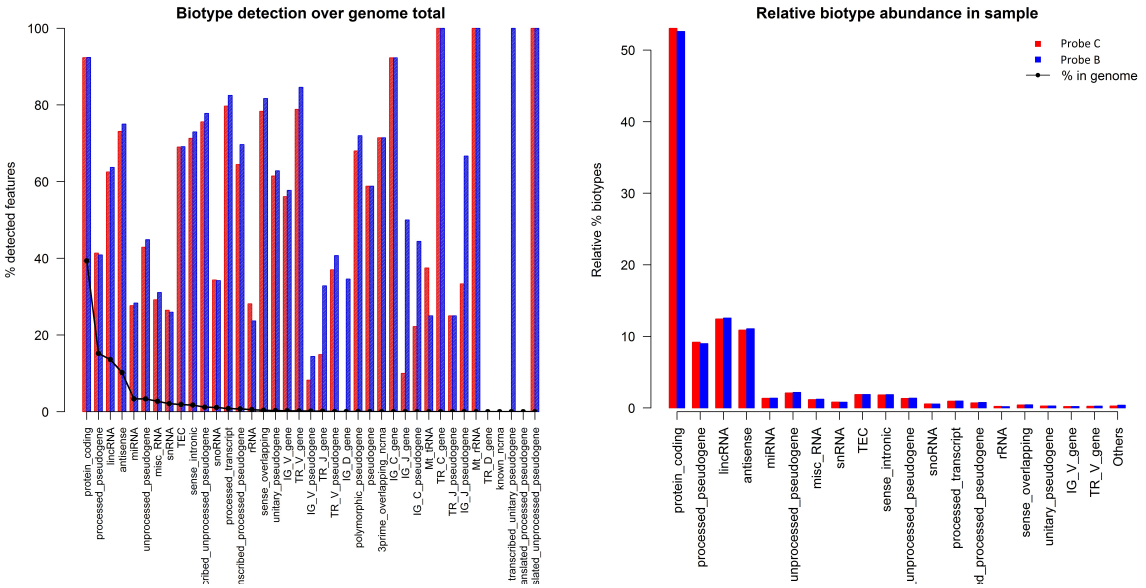
Supplementary Figure A.28: Histograms of the number of genes with at least 1, 5, 10, 50 counts per sample.



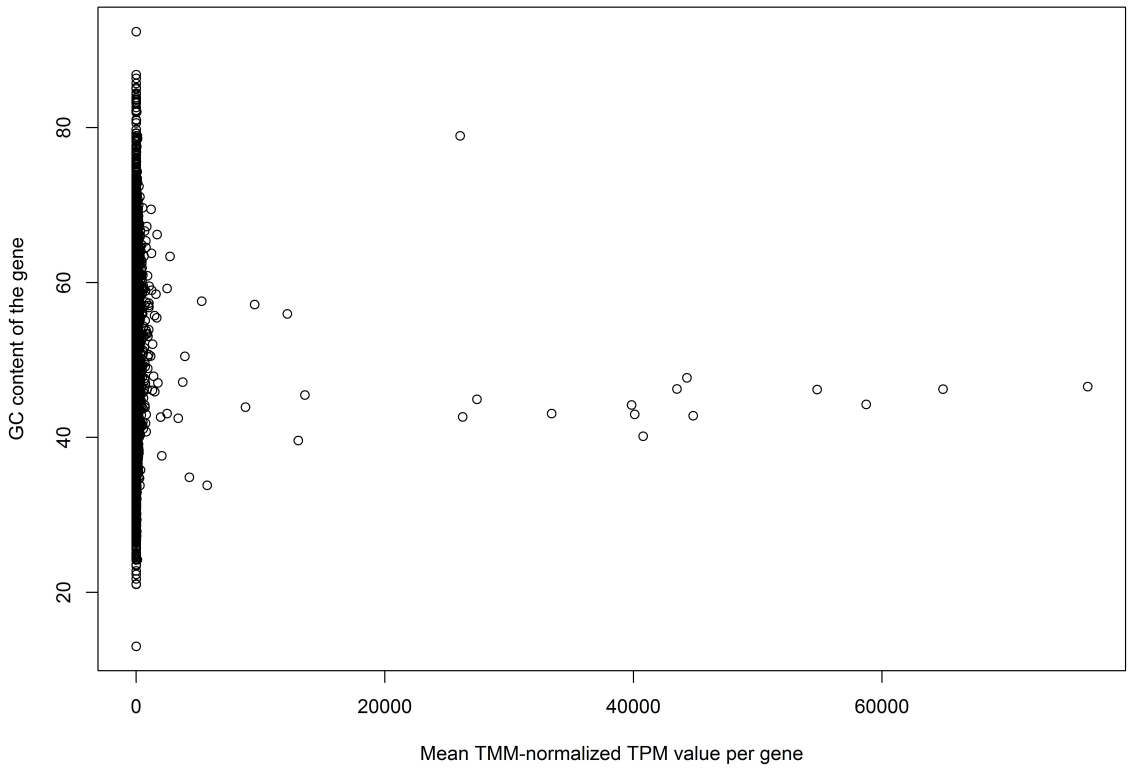
Supplementary Figure A.29: Histogram of the average read counts per gene. The maximum average count of a gene is 457,066 using 6 samples per lane, 992,782 using 3 samples per lane, and 2,995,596 using 1 sample per lane.

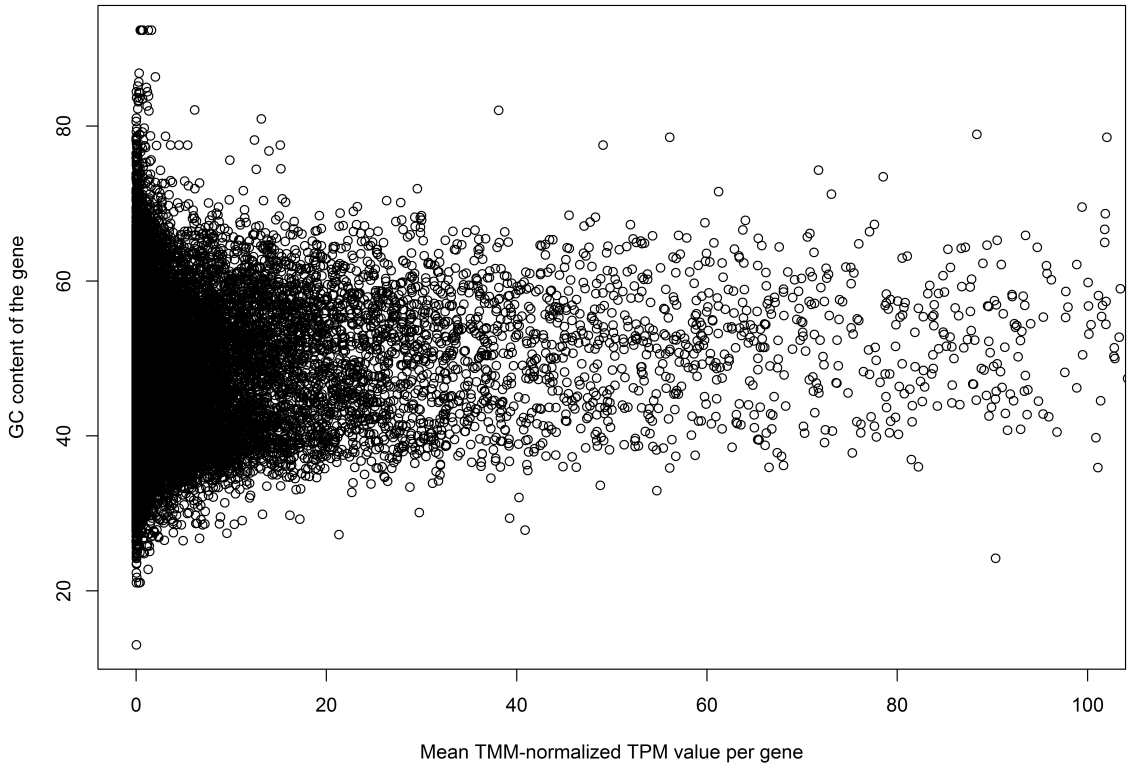


Supplementary Figure A.30: Biodetection plot contrasting the detected features (with at least 1 count) in one deeply sequenced probe ("Probe B") with 1 sample per lane, and one probe ("Probe A") sequenced with 6 probes per lane, generated with the R package `NOISeq`.

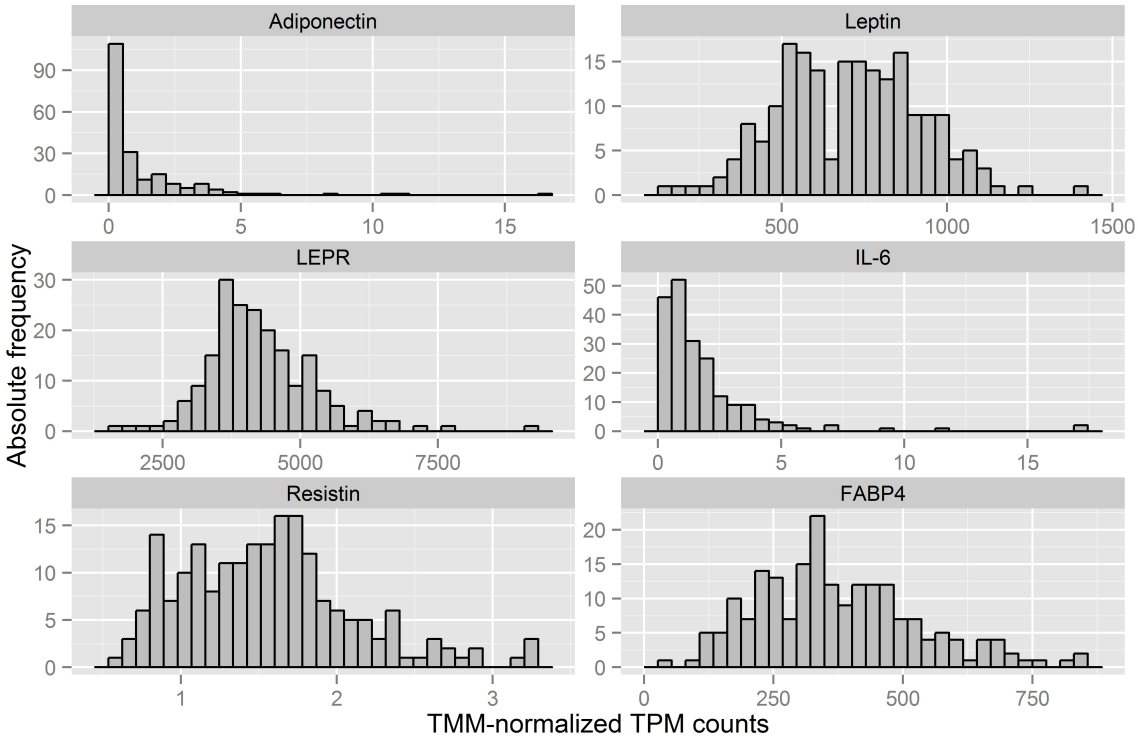


Supplementary Figure A.31: Biodetection plot contrasting the detected features (with at least 1 count) of the two deeply sequenced probe with 1 sample per lane, generated with the R package `NOISeq`.

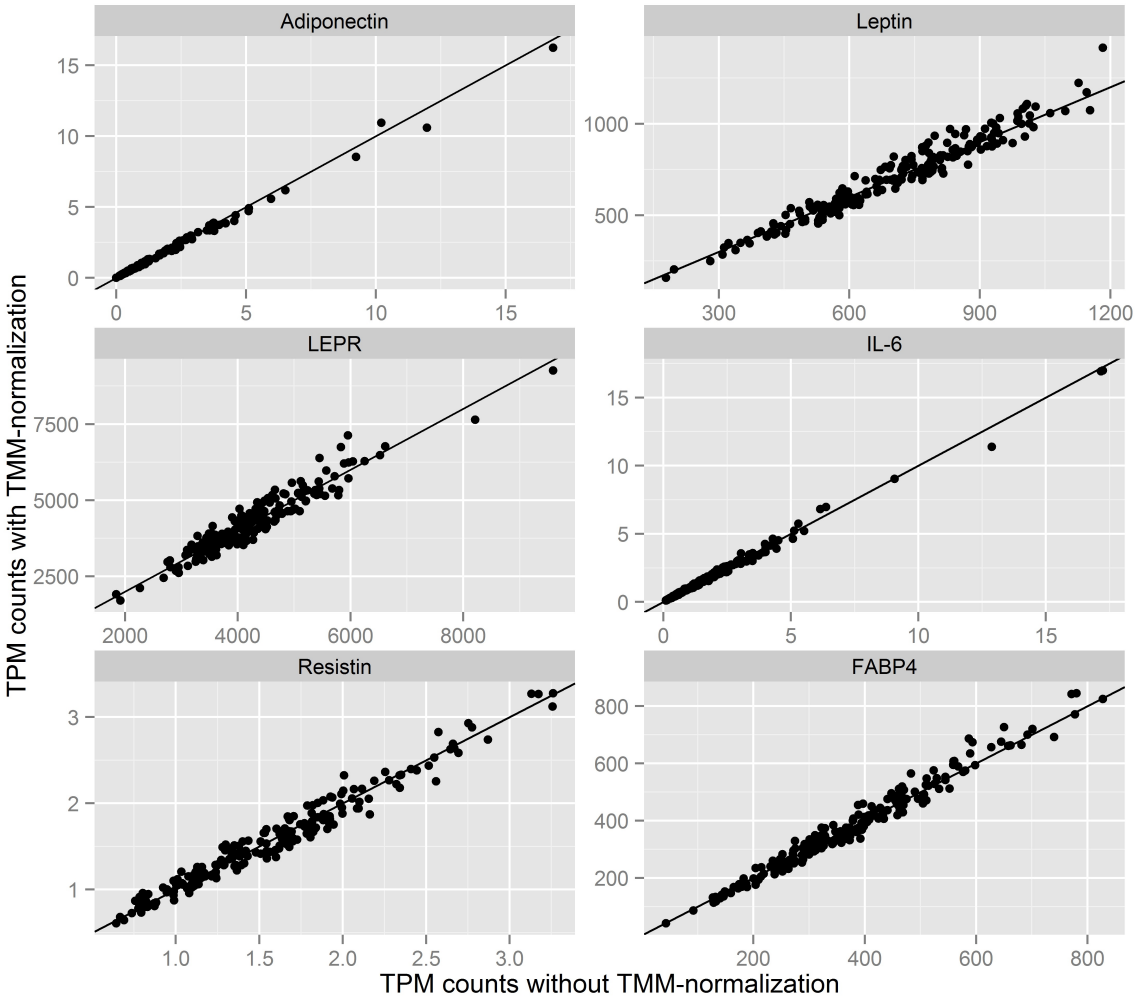




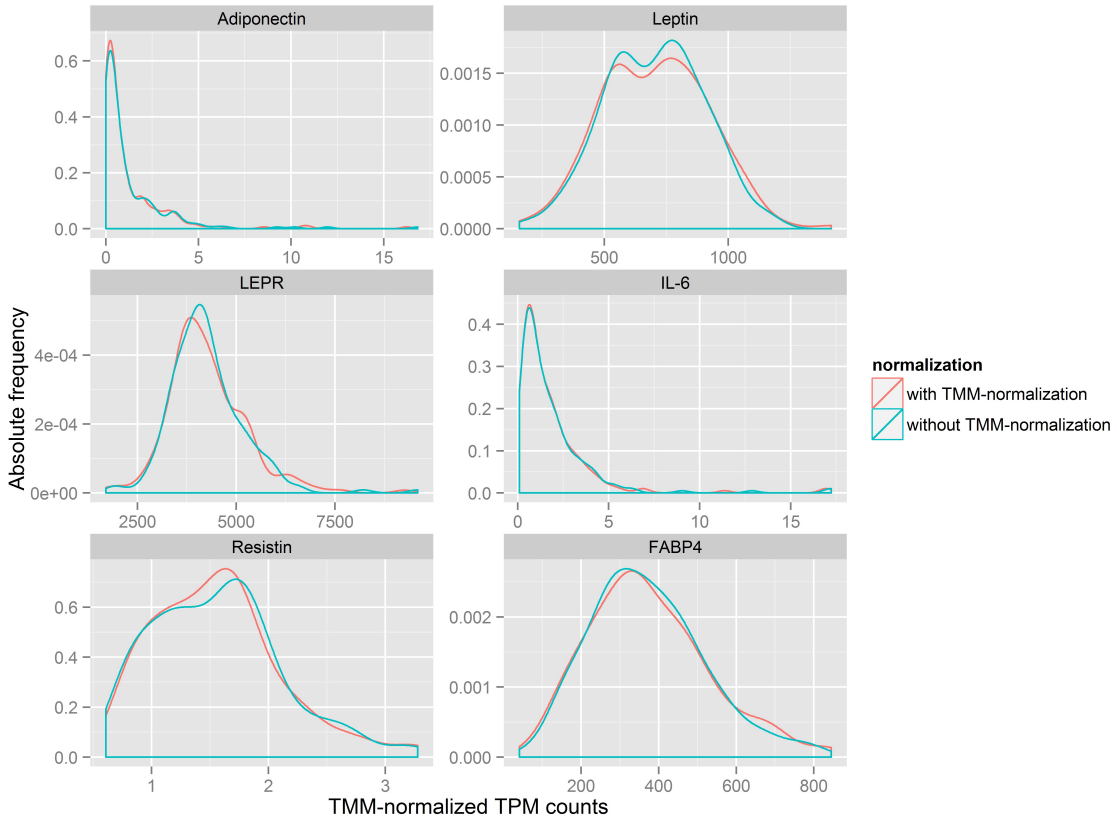
Supplementary Figure A.32: Scatterplots of the mean gene expression (over all samples) of a gene versus its GC content, for all genes (upper panel) and zoomed in to the genes with lower expression (lower panel).



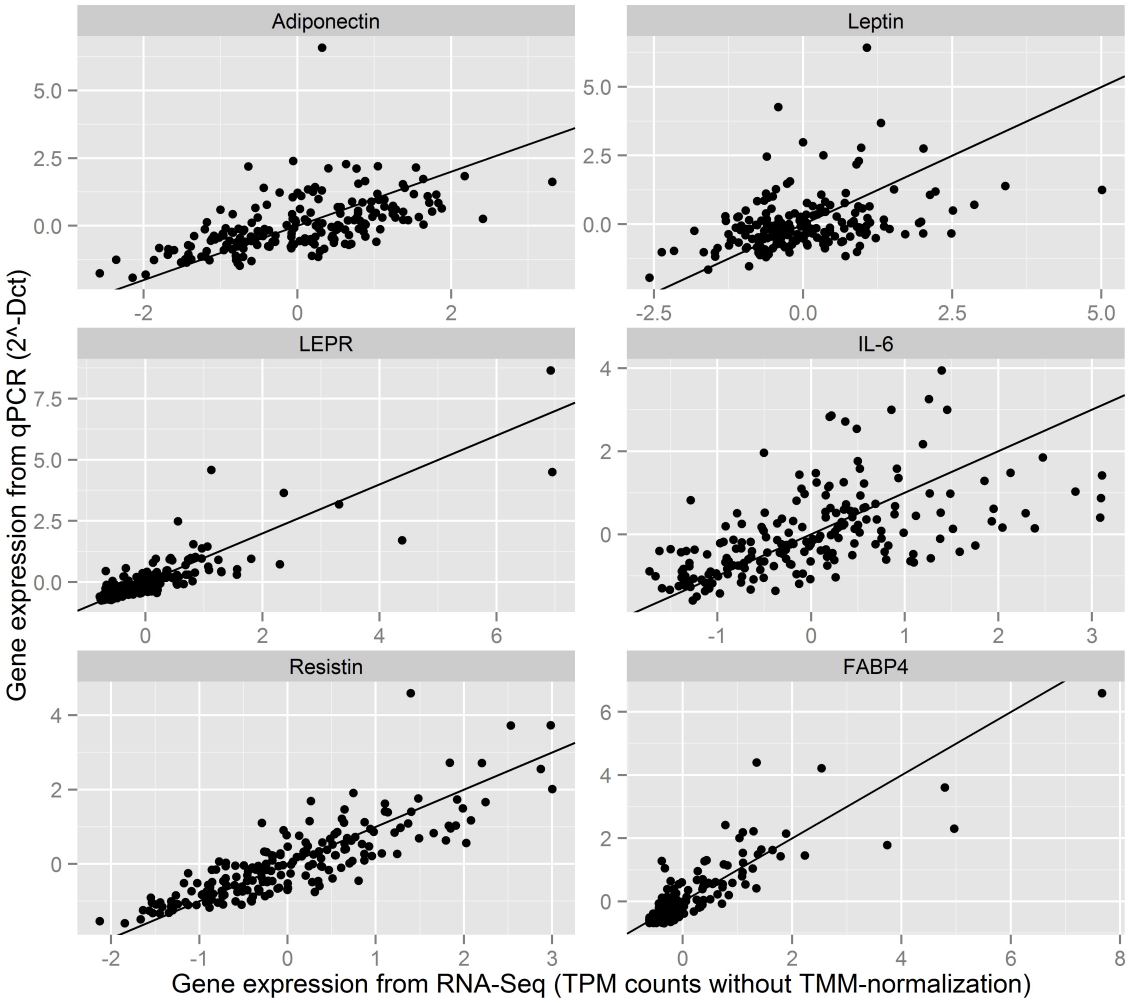
Supplementary Figure A.33: Histograms of the TMM-normalized TPM counts for the 6 candidate genes investigated in the study described in section A.4.5.



Supplementary Figure A.34: Scatterplots of the TPM counts for the 6 candidate genes investigated in the study described in section A.4.5, with and without TMM-normalization, with the diagonal for reference.

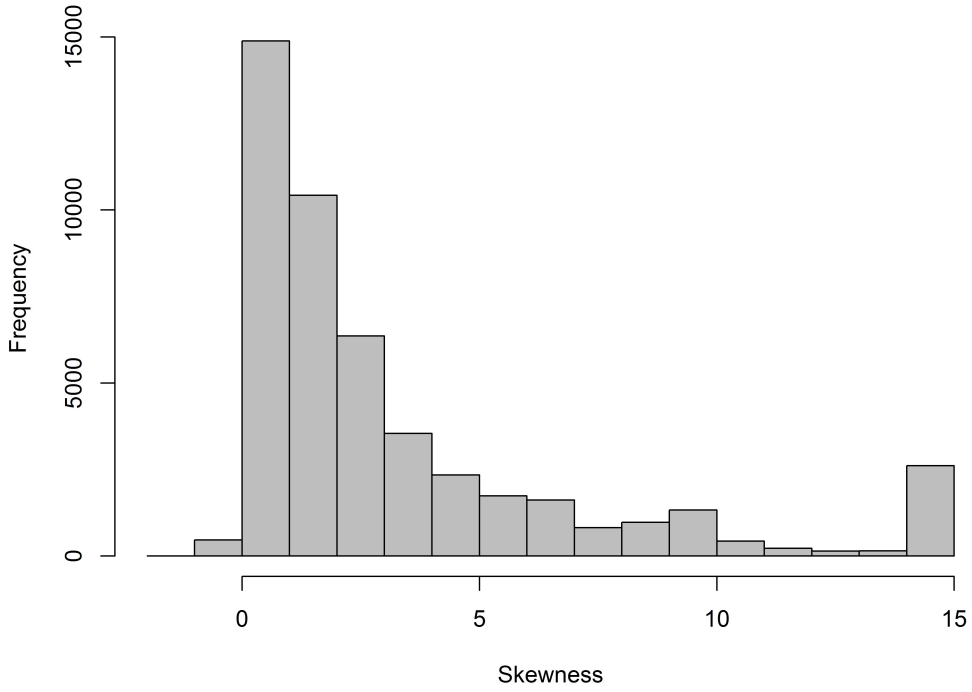


Supplementary Figure A.35: Density estimates of the distribution of the TPM counts for the 6 candidate genes investigated in the study described in section A.4.5, with and without TMM-normalization.

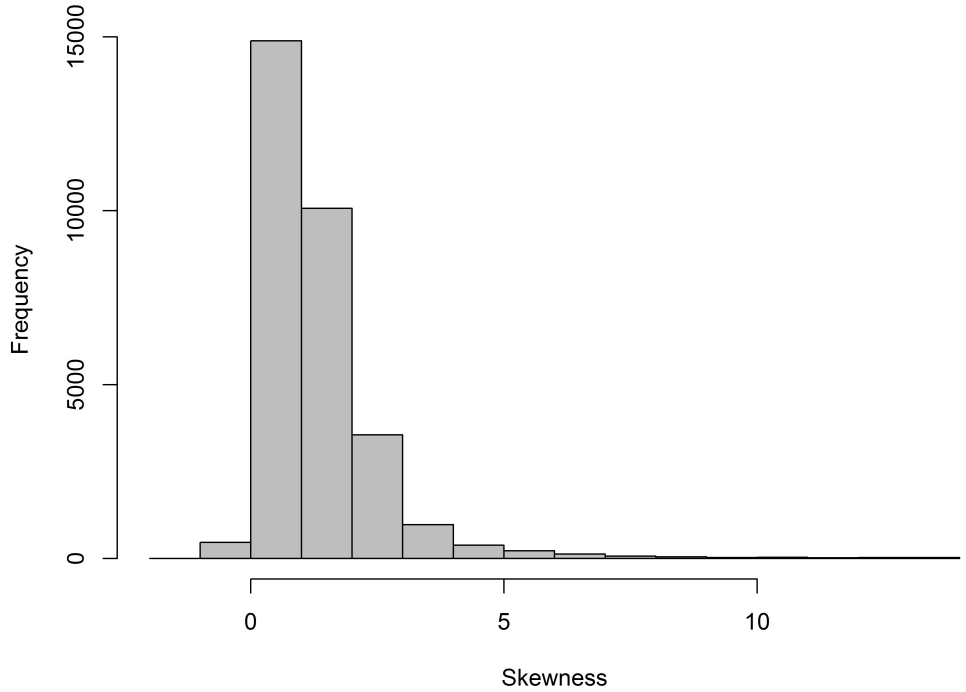


Supplementary Figure A.36: Scatterplots between the gene expression measures from RNA-seq and qPCR for the 6 candidate genes investigated in the study described in section A.4.5, with the diagonal for reference.

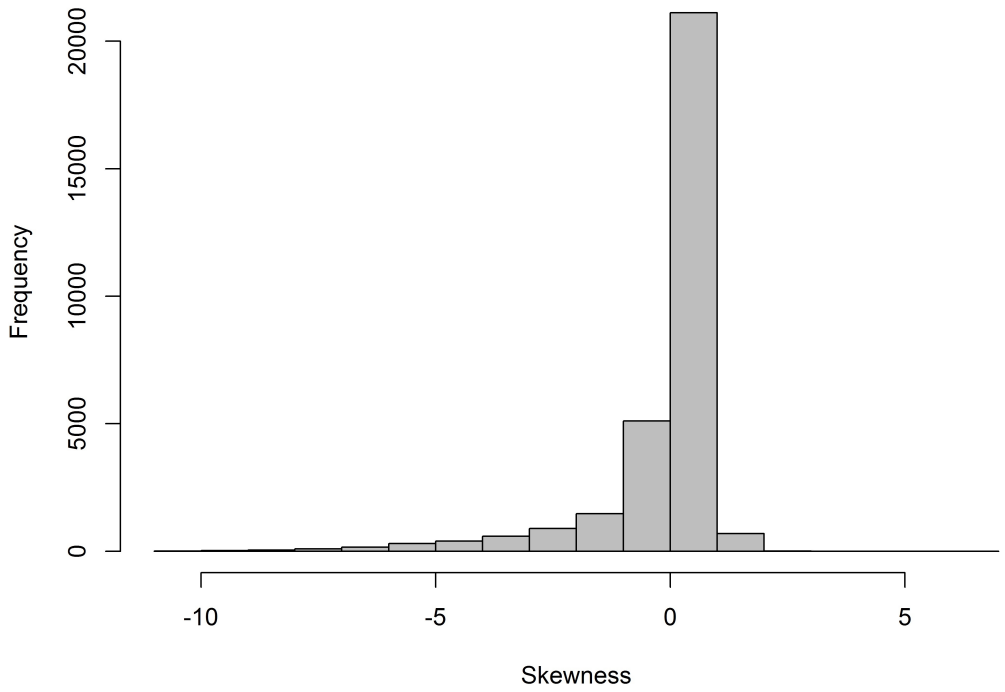
A.6.5.5 Distribution and transformations of normalized read counts



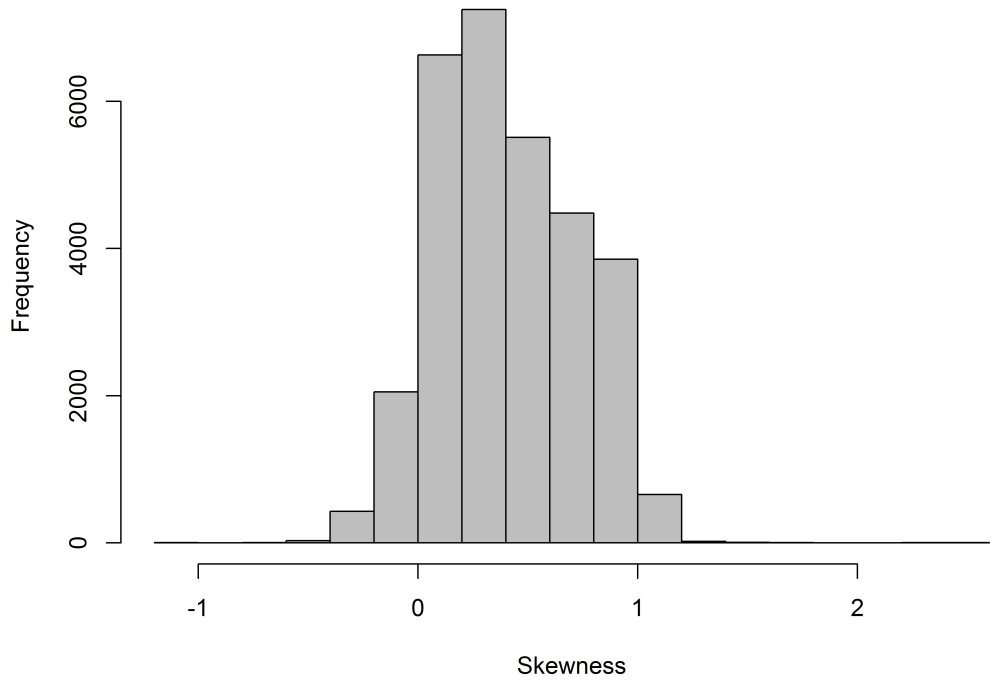
Supplementary Figure A.37: Histogram of the skewness index in all 48,019 expressed genes with non-zero counts for at least 1 sequenced probe.



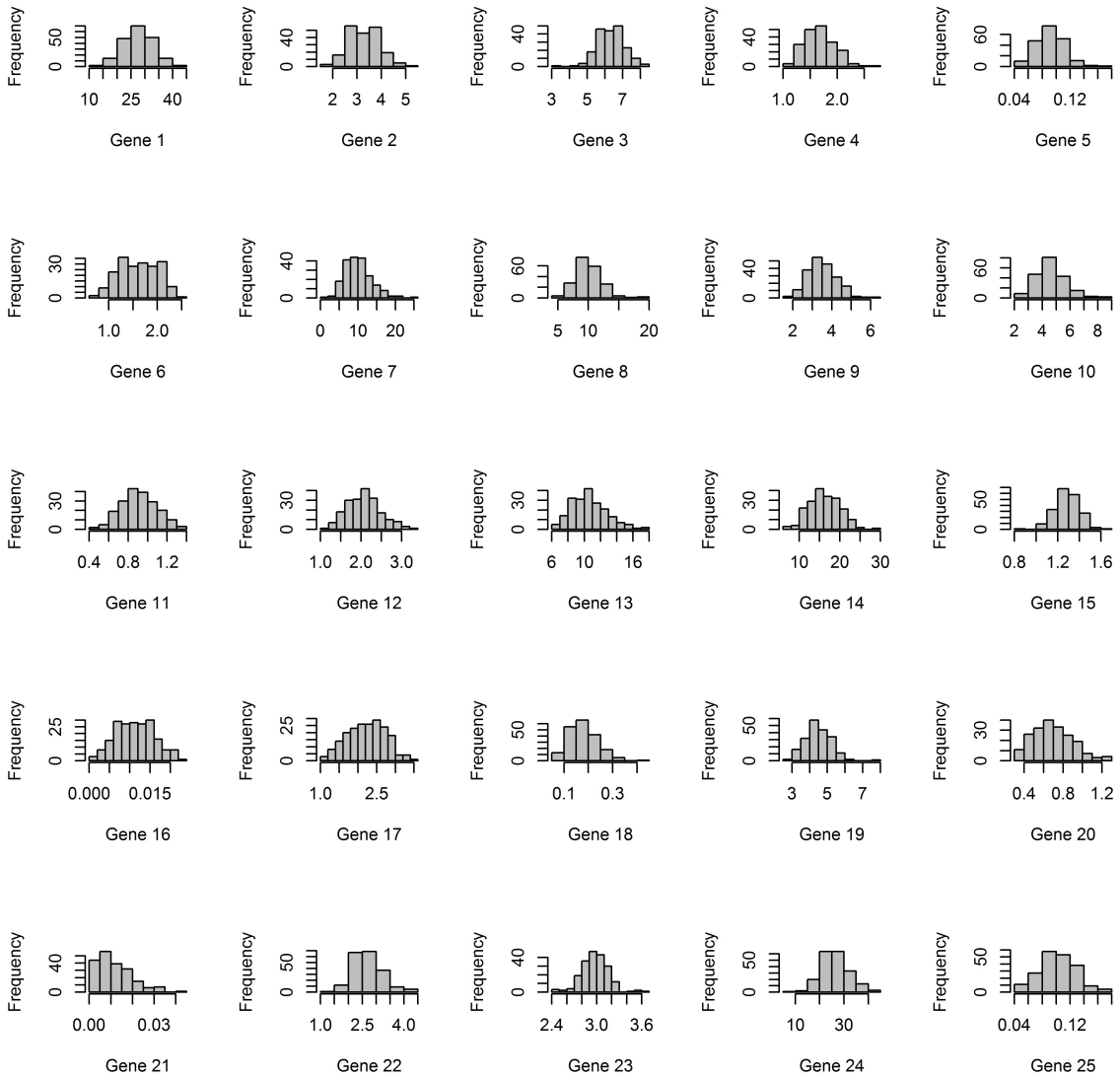
Supplementary Figure A.38: Histogram of the skewness index in the 30,917 genes with non-zero counts for at least 25% percent of all probes.



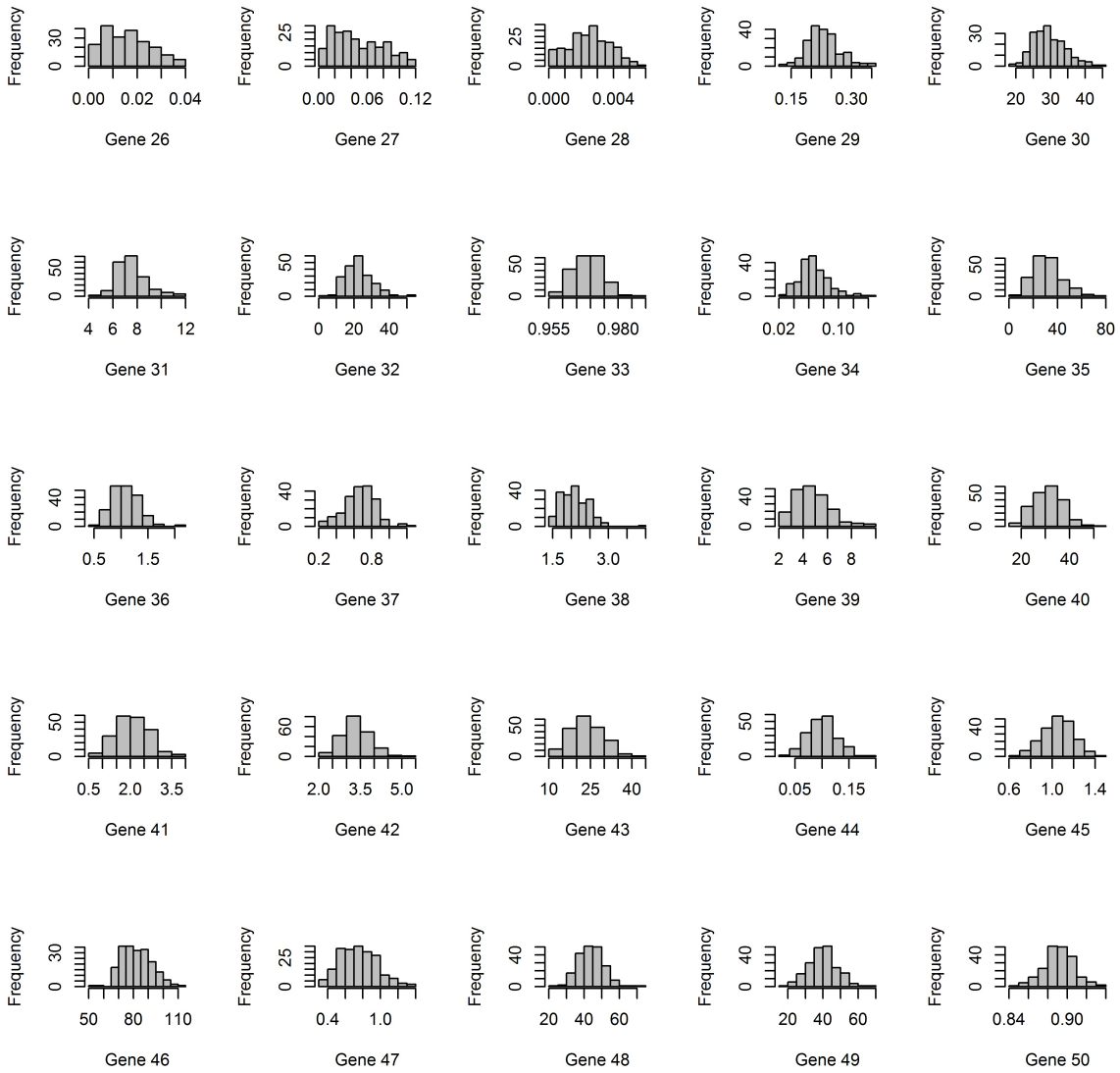
Supplementary Figure A.39: Histogram of the skewness index in the 30,917 genes with non-zero counts for at least 25% percent of all probes, after those genes with skewness greater than 1 were log-transformed (after adding the constant 0.000001).



Supplementary Figure A.40: Histogram of the skewness index in the 30,917 genes with non-zero counts for at least 25% percent of all probes, after those genes with skewness greater than 1 were transformed using the Yeo-Johnson transformation in the `car` R package.

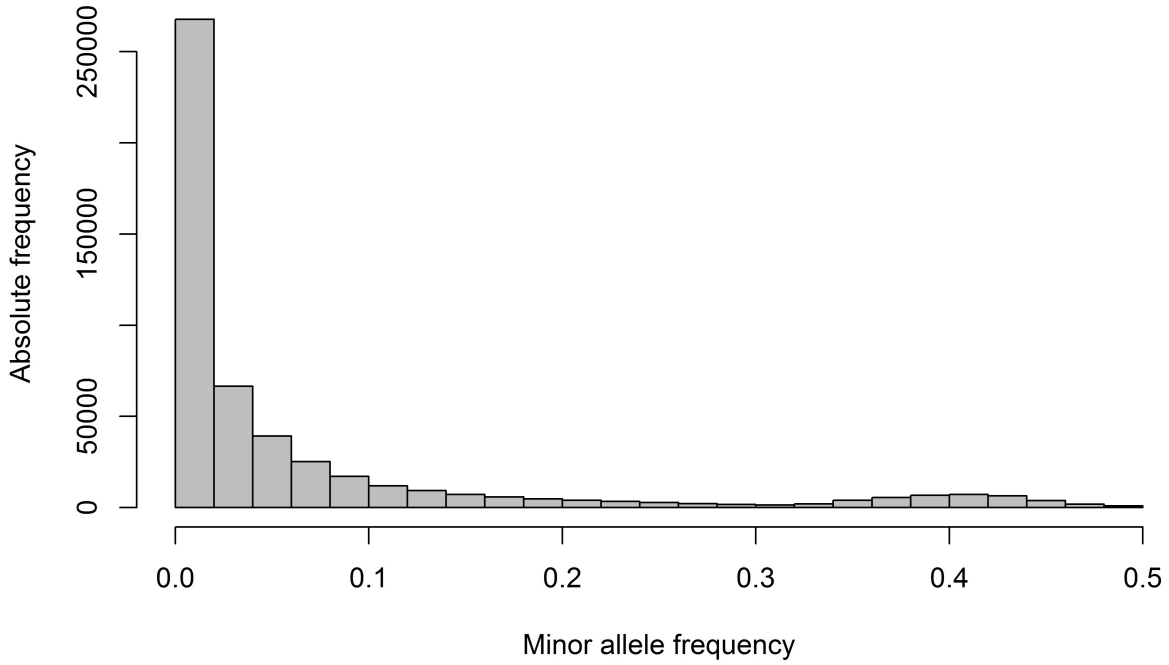


Supplementary Figure A.41: Histograms of the first 25 of the 30,917 genes with non-zero counts for at least 25% percent of all probes, after those genes with skewness greater than 1 were transformed using the Yeo-Johnson transformation in the `car` R package



Supplementary Figure A.42: Histograms of genes 26-50 of the 30,917 genes with non-zero counts for at least 25% percent of all probes, after those genes with skewness greater than 1 were transformed using the Yeo-Johnson transformation in the `car` R package.

A.6.5.6 Descriptive statistics of the called SNVs



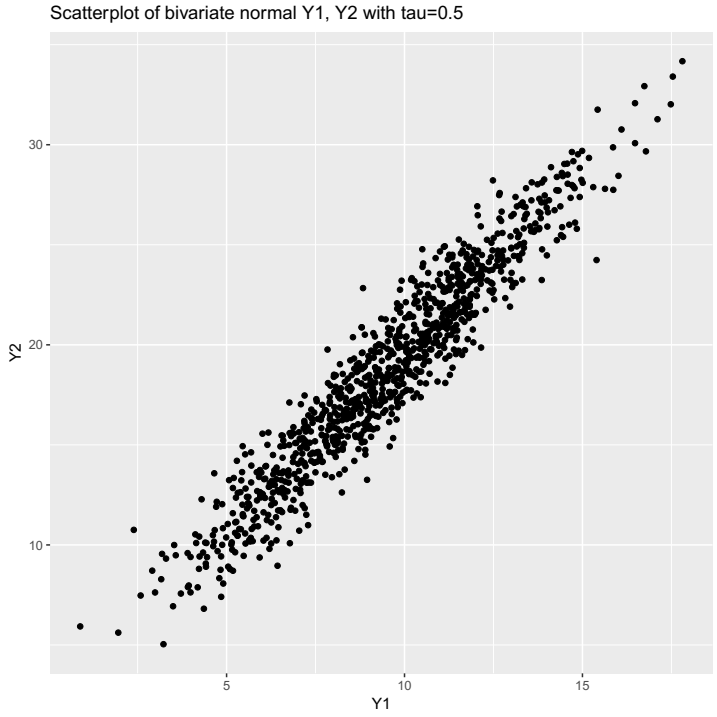
Supplementary Figure A.43: Histogram of the MAF of the quality-controlled 509,009 SNVs that were analyzed in the main analysis of the obesity study.

A.7 R code

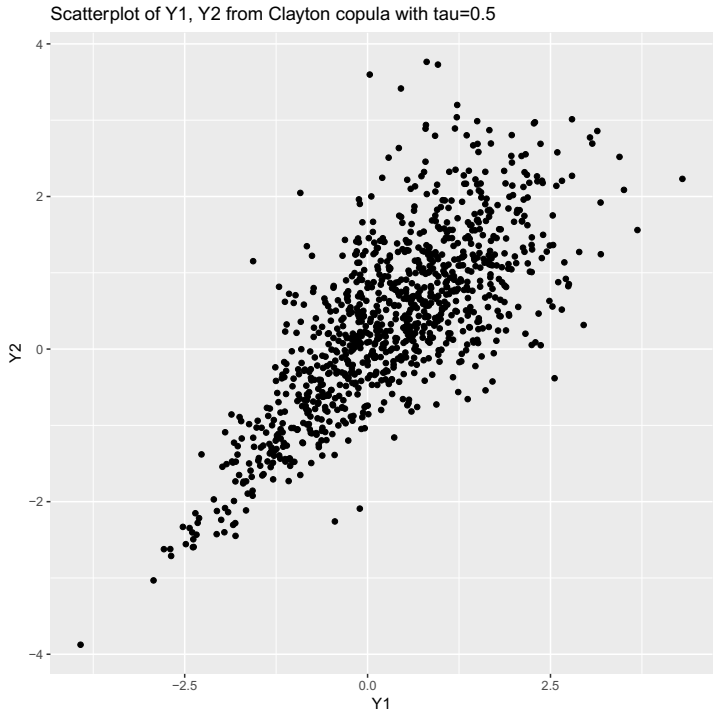
In the following, the R functions of the `CJAMP` and `CIEE` packages are illustrated in their use to generate sample data, fit a copula model, and show the formatted output.

A.7.1 C-JAMP

```
# Load C-JAMP package
library(CJAMP)
library(ggplot2)

# Generate genetic data:
genodata <- generate_genodata(n_SNV = 20, n_ind = 1000)
compute_MAF(genodata)

##  SNV1  SNV2  SNV3  SNV4  SNV5  SNV6  SNV7  SNV8  SNV9 SNV10 SNV11
## 0.357 0.287 0.060 0.456 0.253 0.227 0.476 0.060 0.325 0.116 0.003
## SNV12 SNV13 SNV14 SNV15 SNV16 SNV17 SNV18 SNV19 SNV20
## 0.298 0.150 0.022 0.272 0.180 0.229 0.477 0.155 0.177

# Generate phenotype data from the Clayton copula:
dat1a <- generate_clayton_copula(n = 1000, phi = 0.5)

head(dat1a)

##           Y1          Y2
## 1 -0.4330635  0.69184891
## 2 -0.8775747  0.09145315
## 3 -0.6970692  0.68681678
## 4  0.6685981 -0.27464723
## 5  0.1135054  0.92042670
## 6  0.5343440  0.27794493

dat1b <- generate_clayton_copula(n = 1000, phi = 2)
dat1c <- generate_clayton_copula(n = 1000, phi = 8)

# Generate phenotypes from bivariate normal distribution
# given covariates:
phenodata_bvn <- generate_phenodata_2_bvn(genodata = genodata,
                                          tau = 0.5, b1 = 1, b2 = 2)
ggplot(phenodata_bvn, aes(x=Y1, y=Y2)) + geom_point() +
  labs(title="Scatterplot of bivariate normal Y1, Y2 with tau=0.5")
```




```
# Generate phenotype data from the Clayton copula given covariates:
phenodata <- generate_phenodata_2_copula(genodata = genodata$SNV1,
                                         MAF_cutoff = 1,
                                         prop_causal = 1, tau = 0.5,
                                         b1 = 0.3, b2 = 0.3)

ggplot(phenodata, aes(x=Y1, y=Y2)) + geom_point() +
  labs(title="Scatterplot of Y1, Y2 from Clayton copula with tau=0.5")
```



```r
# Check how much of the phenotypic variance (of Y1) is explained
# by the genetic markers:
compute_expl_var(genodata = genodata, phenodata = phenodata$Y1,
                 type = c("Rsquared_unadj", "Rsquared_adj",
                          "MAF_based", "MAF_based_Y_adjusted"),
                 causal_idx = rep(TRUE,20), effect_causal =
                 c(0.3 * abs(log10(compute_MAF(genodata$SNV1))),
                 rep(0,19)))

## $Rsquared_unadj
## [1] 0.02028775
##
## $Rsquared_adj
## [1] 0.0002731993
##
## $MAF_based
## [1] 0.008268204
##
## $MAF_based_Y_adjusted
## [1] 0.006200331

# Compute naive estimates of all parameters in the model
predictors <- data.frame(X1 = phenodata$X1, X2 = phenodata$X2,
                         SNV = genodata$SNV1)
get_estimates_naive(Y1 = phenodata$Y1, Y2 = phenodata$Y2,
                    predictors_Y1 = predictors,
                    predictors_Y2 = predictors,
                    copula_param = "both")

##          log_phi log_theta_minus1     Y1_log_sigma     Y2_log_sigma
##      0.872799990      0.179652810      0.029125972      0.032939381
##   Y1_(Intercept)            Y1_X1            Y1_X2           Y1_SNV
##      0.007367563      0.459359315      0.530724415      0.079456860
##   Y2_(Intercept)            Y2_X1            Y2_X2           Y2_SNV
##      0.030867828      0.423911195      0.531748193      0.069895600

# Compute the minus log-likelihood of the copula model
predictors <- data.frame(X1 = phenodata$X1, X2 = phenodata$X2,
                         genodata[, 1:5])
estimates <- get_estimates_naive(Y1 = phenodata$Y1,
                                 Y2 = phenodata$Y2,
                                 predictors_Y1 = predictors,
                                 predictors_Y2 = predictors,
                                 copula_param = "both")
```

```
minusloglik(Y1 = phenodata$Y1, Y2 = phenodata$Y2,
            predictors_Y1 = predictors, predictors_Y2 = predictors,
            parameters = estimates, copula = "2param")

## [1] 3245.192

# Perform LRT to test whether 2-parameter copula model has a better
# model fit compared to Clayton copula (no).
predictors <- data.frame(X1 = phenodata$X1, X2 = phenodata$X2,
                         SNV = genodata$SNV1)
estimates_c <- get_estimates_naive(Y1 = phenodata$Y1,
                                   Y2 = phenodata$Y2,
                                   predictors_Y1 = predictors,
                                   predictors_Y2 = predictors,
                                   copula_param = "phi")
minusloglik_Clayton <- minusloglik(Y1 = phenodata$Y1,
                                   Y2 = phenodata$Y2,
                                   predictors_Y1 = predictors,
                                   predictors_Y2 = predictors,
                                   parameters = estimates_c,
                                   copula = "Clayton")
estimates_2p <- get_estimates_naive(Y1 = phenodata$Y1,
                                    Y2 = phenodata$Y2,
                                    predictors_Y1 = predictors,
                                    predictors_Y2 = predictors,
                                    copula_param = "both")
minusloglik_2param <- minusloglik(Y1 = phenodata$Y1,
                                  Y2 = phenodata$Y2,
                                  predictors_Y1 = predictors,
                                  predictors_Y2 = predictors,
                                  parameters = estimates_2p,
                                  copula = "2param")
lrt_copula(minusloglik_Clayton, minusloglik_2param)

## $chisq
## [1] -1552.009
##
## $pval
## [1] 0.5

# Perform LRT to test the marginal parameters
# (alternative model has better fit).
predictors_1 <- data.frame(X1 = phenodata$X1, X2 = phenodata$X2)
estimates_1 <- get_estimates_naive(Y1 = phenodata$Y1,
```

```
                                    Y2 = phenodata$Y2,
                                    predictors_Y1 = predictors_1,
                                    predictors_Y2 = predictors_1,
                                    copula = "phi")
minusloglik_1 <- minusloglik(Y1 = phenodata$Y1, Y2 = phenodata$Y2,
                             predictors_Y1 = predictors_1,
                             predictors_Y2 = predictors_1,
                             parameters = estimates_1,
                             copula = "Clayton")
predictors_2 <- data.frame(X1 = phenodata$X1, X2 = phenodata$X2,
                           SNV = genodata$SNV1)
estimates_2 <- get_estimates_naive(Y1 = phenodata$Y1,
                                    Y2 = phenodata$Y2,
                                    predictors_Y1 = predictors_2,
                                    predictors_Y2 = predictors_2,
                                    copula = "phi")
minusloglik_2 <- minusloglik(Y1 = phenodata$Y1, Y2 = phenodata$Y2,
                             predictors_Y1 = predictors_2,
                             predictors_Y2 = predictors_2,
                             parameters = estimates_2,
                             copula = "Clayton")
lrt_param(minusloglik_1, minusloglik_2, df=2)

## $chisq
## [1] 6.53667
##
## $pval
## [1] 0.03806977

# Main function of the package: Compute C-JAMP and show results
predictors <- data.frame(X1 = phenodata$X1, X2 = phenodata$X2,
                         genodata[, 1:3])
cjamp_res <- cjamp(copula = "2param", Y1 = phenodata$Y1,
                   Y2 = phenodata$Y2, predictors_Y1 = predictors,
                   predictors_Y2 = predictors, scale_var = FALSE,
                   optim_method = "BFGS", trace = 0,
                   kkt2tol = 1E-16, SE_est = TRUE, pval_est = TRUE,
                   n_iter_max = 10)
cjamp_res

## $`Parameter point estimates`
##       Y1_sigma       Y2_sigma Y1_(Intercept)          Y1_X1
##   1.018039e+00   1.024030e+00  -3.263697e-03   4.577113e-01
##          Y1_X2        Y1_SNV1        Y1_SNV2        Y1_SNV3
```

```
##   4.778287e-01   1.111108e-01  -7.478800e-03   1.524211e-01
## Y2_(Intercept)          Y2_X1          Y2_X2        Y2_SNV1
##   3.225616e-03   4.562371e-01   5.021090e-01   1.072456e-01
##        Y2_SNV2        Y2_SNV3            phi          theta
##  -3.026763e-03   1.119920e-01   2.029883e+00   1.000002e+00
##            tau       lambda_l       lambda_u
##   5.037089e-01   7.107243e-01   3.417528e-06
##
## $`Parameter standard error estimates`
##       Y1_sigma       Y2_sigma Y1_(Intercept)          Y1_X1
##   0.0207191670   0.0208030635   0.0595696133   0.0274704660
##         Y1_X2        Y1_SNV1        Y1_SNV2        Y1_SNV3
##   0.0560816396   0.0412244792   0.0434147479   0.0814883054
## Y2_(Intercept)          Y2_X1          Y2_X2        Y2_SNV1
##   0.0602285864   0.0278789350   0.0562067891   0.0418147801
##        Y2_SNV2        Y2_SNV3            phi          theta
##   0.0435238260   0.0829850914   0.1179059134   0.0004222667
##            tau       lambda_l       lambda_u
##   0.0145197170   0.0140960373   0.0005853820
##
## $`Parameter p-values`
## Y1_(Intercept)          Y1_X1          Y1_X2        Y1_SNV1
##   9.563074e-01   2.478419e-62   1.592311e-17   7.033319e-03
##        Y1_SNV2        Y1_SNV3 Y2_(Intercept)          Y2_X1
##   8.632300e-01   6.141915e-02   9.572887e-01   3.403589e-60
##         Y2_X2        Y2_SNV1        Y2_SNV2        Y2_SNV3
##   4.136946e-19   1.032421e-02   9.445577e-01   1.771624e-01
##
## $`Convergence code of optimx function`
## convcode
##        0
##
## $`Karush-Kuhn-Tucker conditions 1 and 2`
## KKT1 KKT2
## TRUE TRUE
##
## $`Maximum log-likelihood`
## maxloglik
##  2443.347
##
## attr(,"class")
## [1] "cjamp"

# Wrapper of cjamp function to analyze multiple SNVs sequentially
```

```
covariates <- data.frame(X1 = phenodata$X1, X2 = phenodata$X2)
predictors <- genodata
cjamp_loop_res <- cjamp_loop(copula = "Clayton", Y1 = phenodata$Y1,
                             Y2 = phenodata$Y2,
                             predictors = predictors,
                             covariates_Y1 = covariates,
                             covariates_Y2 = covariates,
                             scale_var = FALSE,
                             optim_method = "BFGS",
                             trace = 0, kkt2tol = 1E-16,
                             SE_est = TRUE, pval_est = TRUE,
                             n_iter_max = 10)
cjamp_loop_res
```

```
## $`Parameter point estimates for effects of predictors on Y1`
##         SNV1         SNV2         SNV3         SNV4         SNV5
##  0.106697023 -0.006888141  0.140730286  0.043119631 -0.056822154
##         SNV6         SNV7         SNV8         SNV9        SNV10
##  0.061547846  0.042328067 -0.028755051  0.012920556 -0.029087665
##        SNV11        SNV12        SNV13        SNV14        SNV15
##  0.399968842  0.008051045  0.053995839  0.054145340  0.038479840
##        SNV16        SNV17        SNV18        SNV19        SNV20
## -0.038479431  0.029074078 -0.028184904 -0.023865019  0.003464513
##
## $`Parameter point estimates for effects of predictors on Y2`
##         SNV1         SNV2         SNV3         SNV4         SNV5
##  0.104571428 -0.002372445  0.102154014  0.003258226 -0.055813884
##         SNV6         SNV7         SNV8         SNV9        SNV10
##  0.055261870  0.033801549  0.007901566  0.030418373  0.093229315
##        SNV11        SNV12        SNV13        SNV14        SNV15
##  0.031250816  0.010929721  0.099484751  0.088670269  0.029236045
##        SNV16        SNV17        SNV18        SNV19        SNV20
## -0.005803029  0.053461354 -0.041143949 -0.035529587  0.041571513
##
## $`Parameter standard error estimates for effects on Y1`
##       SNV1       SNV2       SNV3       SNV4       SNV5       SNV6
## 0.04114480 0.04360486 0.08148113 0.03909485 0.04596326 0.04707007
##       SNV7       SNV8       SNV9      SNV10      SNV11      SNV12
## 0.03940785 0.08367532 0.04212457 0.06268610 0.37252599 0.04304969
##      SNV13      SNV14      SNV15      SNV16      SNV17      SNV18
## 0.05582313 0.14138213 0.04399573 0.05163744 0.04726034 0.03981132
##      SNV19      SNV20
## 0.05493949 0.05122328
##
```

```
## $`Parameter standard error estimates for effects on Y2`
##       SNV1       SNV2       SNV3       SNV4       SNV5       SNV6
## 0.04180830 0.04373260 0.08309615 0.03978259 0.04602636 0.04735030
##       SNV7       SNV8       SNV9      SNV10      SNV11      SNV12
## 0.04000655 0.08412875 0.04174953 0.06280148 0.38128497 0.04401251
##      SNV13      SNV14      SNV15      SNV16      SNV17      SNV18
## 0.05628767 0.14468444 0.04405795 0.05181463 0.04736861 0.03963240
##      SNV19      SNV20
## 0.05466032 0.05170494
##
## $`Parameter p-values for effects of predictors on Y1`
##        SNV1        SNV2        SNV3        SNV4        SNV5
## 0.009508533 0.874482590 0.084140406 0.270049290 0.216365020
##        SNV6        SNV7        SNV8        SNV9       SNV10
## 0.191016022 0.282776731 0.731109263 0.759054548 0.642632736
##       SNV11       SNV12       SNV13       SNV14       SNV15
## 0.282971981 0.851646923 0.333410825 0.701740823 0.381777009
##       SNV16       SNV17       SNV18       SNV19       SNV20
## 0.456160068 0.538429355 0.478968782 0.664007179 0.946075777
##
## $`Parameter p-values for effects of predictors on Y2`
##       SNV1       SNV2       SNV3       SNV4       SNV5       SNV6
## 0.01237691 0.95673686 0.21894168 0.93472560 0.22526351 0.24317562
##       SNV7       SNV8       SNV9      SNV10      SNV11      SNV12
## 0.39816642 0.92517087 0.46625126 0.13767413 0.93467706 0.80387743
##      SNV13      SNV14      SNV15      SNV16      SNV17      SNV18
## 0.07715553 0.53997365 0.50695814 0.91082662 0.25905640 0.29920523
##      SNV19      SNV20
## 0.51568773 0.42138871
##
## $`Convergence code of optimx function`
##  [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##
## $`Karush-Kuhn-Tucker conditions 1 and 2`
##  [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [12] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##
## $`Maximum log-likelihood`
##  [1] 2445.109 2448.779 2447.307 2447.575 2447.934 2447.893 2448.249
##  [8] 2448.587 2448.464 2444.984 2447.911 2448.766 2447.075 2448.593
## [15] 2448.414 2448.286 2448.095 2448.245 2448.581 2448.143
##
## attr(,"class")
```

```
## [1] "cjamp"

# Summary of regular cjamp function output
summary(cjamp_res)

## [1] "C-JAMP estimates of marginal parameters."
##                point_estimates SE_estimates      pvalues
## Y1_(Intercept)    -0.003263697   0.05956961 9.563074e-01
## Y1_X1              0.457711329   0.02747047 2.478419e-62
## Y1_X2              0.477828702   0.05608164 1.592311e-17
## Y1_SNV1            0.111110794   0.04122448 7.033319e-03
## Y1_SNV2           -0.007478800   0.04341475 8.632300e-01
## Y1_SNV3            0.152421095   0.08148831 6.141915e-02
## Y2_(Intercept)     0.003225616   0.06022859 9.572887e-01
## Y2_X1              0.456237118   0.02787893 3.403589e-60
## Y2_X2              0.502108981   0.05620679 4.136946e-19
## Y2_SNV1            0.107245591   0.04181478 1.032421e-02
## Y2_SNV2           -0.003026763   0.04352383 9.445577e-01
## Y2_SNV3            0.111992006   0.08298509 1.771624e-01

# Summary of looped cjamp function output
summary(cjamp_loop_res)

## [1] "C-JAMP estimates of marginal parameters on Y1."
##          point_estimates SE_estimates     pvalues
## Y1_SNV1      0.106697023   0.04114480 0.009508533
## Y1_SNV2     -0.006888141   0.04360486 0.874482590
## Y1_SNV3      0.140730286   0.08148113 0.084140406
## Y1_SNV4      0.043119631   0.03909485 0.270049290
## Y1_SNV5     -0.056822154   0.04596326 0.216365020
## Y1_SNV6      0.061547846   0.04707007 0.191016022
## Y1_SNV7      0.042328067   0.03940785 0.282776731
## Y1_SNV8     -0.028755051   0.08367532 0.731109263
## Y1_SNV9      0.012920556   0.04212457 0.759054548
## Y1_SNV10    -0.029087665   0.06268610 0.642632736
## Y1_SNV11     0.399968842   0.37252599 0.282971981
## Y1_SNV12     0.008051045   0.04304969 0.851646923
## Y1_SNV13     0.053995839   0.05582313 0.333410825
## Y1_SNV14     0.054145340   0.14138213 0.701740823
## Y1_SNV15     0.038479840   0.04399573 0.381777009
## Y1_SNV16    -0.038479431   0.05163744 0.456160068
## Y1_SNV17     0.029074078   0.04726034 0.538429355
## Y1_SNV18    -0.028184904   0.03981132 0.478968782
## Y1_SNV19    -0.023865019   0.05493949 0.664007179
```

```
## Y1_SNV20     0.003464513   0.05122328 0.946075777
## [1] "C-JAMP estimates of marginal parameters on Y2."
##          point_estimates SE_estimates    pvalues
## Y2_SNV1      0.104571428   0.04180830 0.01237691
## Y2_SNV2     -0.002372445   0.04373260 0.95673686
## Y2_SNV3      0.102154014   0.08309615 0.21894168
## Y2_SNV4      0.003258226   0.03978259 0.93472560
## Y2_SNV5     -0.055813884   0.04602636 0.22526351
## Y2_SNV6      0.055261870   0.04735030 0.24317562
## Y2_SNV7      0.033801549   0.04000655 0.39816642
## Y2_SNV8      0.007901566   0.08412875 0.92517087
## Y2_SNV9      0.030418373   0.04174953 0.46625126
## Y2_SNV10     0.093229315   0.06280148 0.13767413
## Y2_SNV11     0.031250816   0.38128497 0.93467706
## Y2_SNV12     0.010929721   0.04401251 0.80387743
## Y2_SNV13     0.099484751   0.05628767 0.07715553
## Y2_SNV14     0.088670269   0.14468444 0.53997365
## Y2_SNV15     0.029236045   0.04405795 0.50695814
## Y2_SNV16    -0.005803029   0.05181463 0.91082662
## Y2_SNV17     0.053461354   0.04736861 0.25905640
## Y2_SNV18    -0.041143949   0.03963240 0.29920523
## Y2_SNV19    -0.035529587   0.05466032 0.51568773
## Y2_SNV20     0.041571513   0.05170494 0.42138871
```

A.7.2 CIEE

```
# Load CIEE package
library(CIEE)

# Generate sample data with a quantitative primary trait Y
dat <- generate_data(setting="GLM", n = 1000, maf = 0.2,
                     cens = 0.3,a = NULL, b = NULL, aUL = 0,
                     aXL = 0, aXK = 0.2, aLK = 0, aUY = 0,
                     aKY = 0.3, aXY = 0.1, aLY = 0, mu_U = 0,
                     sd_U = 1, X_orth_U = TRUE, mu_X = NULL,
                     sd_X = NULL, mu_L = 0, sd_L = 1, mu_K = 0,
                     sd_K = 1, mu_Y = 0, sd_Y = 1)
head(dat)

##            Y           K X         L           U
## 1  2.4983614 -0.35718347 1  1.298456  0.58709350
## 2 -0.3869956 -0.64368400 1  1.348275  0.03172375
## 3 -2.8471634 -0.61360554 0 -2.221457 -1.39594743
```

```
## 4 -0.3011392 -0.05166774 0 -2.476211  0.26689396
## 5  1.5321377 -0.23907867 0 -1.521010 -0.14122163
## 6 -0.5806856  0.18039008 0  1.133556  0.18819211

# Obtain estimates and perform hypothesis tests under the
# traditional regression approaches
mult_reg(setting = "GLM", Y = dat$Y, X = dat$X, K = dat$K,
         L = dat$L)

## $point_estimates
##     alpha_0     alpha_1    alpha_XY     alpha_2
## -0.01822490  0.27671667  0.13346492 -0.01232202
##
## $SE_estimates
##    alpha_0    alpha_1   alpha_XY    alpha_2
## 0.03687934 0.02978589 0.05247700 0.02960048
##
## $pvalues
##      alpha_0      alpha_1     alpha_XY      alpha_2
## 6.212906e-01 9.406853e-20 1.113124e-02 6.772964e-01

res_reg(Y = dat$Y, X = dat$X, K = dat$K, L = dat$L)

## $point_estimates
##     alpha_0     alpha_1     alpha_2     alpha_3    alpha_XY
##  0.03444987  0.28261304 -0.01268054 -0.05279386  0.13264790
##
## $SE_estimates
##    alpha_0    alpha_1    alpha_2    alpha_3   alpha_XY
## 0.03059805 0.02977685 0.02968121 0.03680861 0.05226472
##
## $pvalues
##      alpha_0      alpha_1      alpha_2      alpha_3     alpha_XY
## 2.604853e-01 1.628744e-20 6.693082e-01 1.518056e-01 1.129989e-02

# Obtain estimates and perform hypothesis tests under the SEM
sem_appl(Y = dat$Y, X = dat$X, K = dat$K, L = dat$L)

## $point_estimates
##     alpha_1     alpha_3     alpha_4     alpha_6    alpha_XY
## -0.01411334  0.13713065 -0.04066894  0.27722407  0.13356893
##
## $SE_estimates
##    alpha_1    alpha_3    alpha_4    alpha_6   alpha_XY
## 0.05593721 0.05554422 0.03139961 0.02970394 0.05237590
```

```
##
## $pvalues
##      alpha_1      alpha_3      alpha_4      alpha_6     alpha_XY
## 8.008039e-01 1.355460e-02 1.952495e-01 1.030064e-20 1.076616e-02

# For CIEE:
# Compute expressions of the functions in equations 3.2.5, 3.2.6
estfunct_GLM <- est_funct_expr(setting="GLM")
estfunct_AFT <- est_funct_expr(setting = "AFT")
estfunct_GLM

## $logL1
## expression(log((1/sqrt(sigma1sq)) * dnorm((y_i - alpha0 - alpha1 *
##     k_i - alpha2 * x_i - alpha3 * l_i)/sqrt(sigma1sq), mean = 0,
##     sd = 1)))
##
## $logL2
## expression(log((1/sqrt(sigma2sq)) * dnorm((y_i - y_bar - alpha1 *
##     (k_i - k_bar) - alpha4 - alphaXY * x_i)/sqrt(sigma2sq), mean = 0,
##     sd = 1)))

estfunct_AFT

## $logL1
## expression(-c_i * log(sigma1) + c_i * log(dnorm((y_i - alpha0 -
##     alpha1 * k_i - alpha2 * x_i - alpha3 * l_i)/sigma1, mean = 0,
##     sd = 1)) + (1 - c_i) * log(1 - pnorm((y_i - alpha0 - alpha1 *
##     k_i - alpha2 * x_i - alpha3 * l_i)/sigma1, mean = 0, sd = 1)))
##
## $logL2
## expression(log((1/sqrt(sigma2sq)) * dnorm(((c_i * y_i + (1 -
##     c_i) * ((alpha0 + alpha1 * k_i + alpha2 * x_i + alpha3 *
##     l_i) + (sigma1 * dnorm((y_i - alpha0 - alpha1 * k_i - alpha2 *
##     x_i - alpha3 * l_i)/sigma1, mean = 0, sd = 1)/(1 - pnorm((y_i -
##     alpha0 - alpha1 * k_i - alpha2 * x_i - alpha3 * l_i)/sigma1,
##     mean = 0, sd = 1))))) - y_adj_bar - alpha1 * (k_i - k_bar) -
##     alpha4 - alphaXY * x_i)/sqrt(sigma2sq), mean = 0, sd = 1)))

# Obtain parameter estimates
estimates <- get_estimates(setting = "GLM", Y = dat$Y, X = dat$X,
                           K = dat$K, L = dat$L)

# Compute the score and hessian matrix
derivobj <- deriv_obj(setting = "GLM", logL1 = estfunct_GLM$logL1,
                      logL2 = estfunct_GLM$logL2, Y = dat$Y, X = dat$X,
```

```
                      K = dat$K, L = dat$L, estimates = estimates)
score_matrix <- scores(derivobj)
hessian_matrix <- hessian(derivobj)

# Compute sandwich standard errors
sandwich_se(scores = score_matrix, hessian = hessian_matrix)

##    alpha_0    alpha_1    alpha_2    alpha_3 sigma_1_sq    alpha_4
## 0.03751627 0.03147702 0.05078096 0.02879671 0.04157306 0.03745880
##   alpha_XY sigma_2_sq
## 0.05076842 0.04155998

# Compute bootstrap standard errors
bootstrap_se(setting = "GLM", BS_rep = 1000, Y = dat$Y, X = dat$X,
             K = dat$K, L = dat$L)

##    alpha_0    alpha_1    alpha_2    alpha_3 sigma_1_sq    alpha_4
## 0.03684819 0.03183278 0.05086000 0.02954175 0.04266883 0.02020396
##   alpha_XY sigma_2_sq
## 0.05076752 0.04264079

# Compute naive standard errors
naive_se(setting = "GLM", Y = dat$Y, X = dat$X, K = dat$K,
         L = dat$L)

##    alpha_0    alpha_1    alpha_2    alpha_3 sigma_1_sq    alpha_4
## 0.03687934 0.02978589 0.05247700 0.02960048         NA 0.03681109
##   alpha_XY sigma_2_sq
## 0.05226824         NA

# Main function to compute CIEE based on all above functions
# Estimates and hypothesis tests under the alternative methods
# can also be performed
results_ciee <- ciee(setting = "GLM", Y = dat$Y, X = dat$X,
                     K = dat$K, L = dat$L, estimates = c("ee",
                     "mult_reg", "res_reg", "sem"), ee_se = "sandwich")
results_ciee

## $results_ee
## $results_ee$point_estimates
##     alpha_0     alpha_1     alpha_2     alpha_3  sigma_1_sq
## -0.01822490  0.27671667  0.13346492 -0.01232202  0.92029662
##     alpha_4    alpha_XY  sigma_2_sq
## -0.05318825  0.13363882  0.92045700
##
```

```
## $results_ee$SE_estimates
##    alpha_0    alpha_1    alpha_2    alpha_3 sigma_1_sq    alpha_4
## 0.03751627 0.03147702 0.05078096 0.02879671 0.04157306 0.03745880
##   alpha_XY sigma_2_sq
## 0.05076842 0.04155998
##
## $results_ee$wald_test_stat
##    alpha_0    alpha_1    alpha_2    alpha_3 sigma_1_sq    alpha_4
## -0.4857865  8.7910704  2.6282473 -0.4278968 22.1368523 -1.4199133
##   alpha_XY sigma_2_sq
##  2.6323219 22.1476747
##
## $results_ee$pvalues
##       alpha_0       alpha_1       alpha_2       alpha_3
##  6.271186e-01  1.481417e-18  8.582611e-03  6.687262e-01
##    sigma_1_sq       alpha_4      alpha_XY    sigma_2_sq
## 1.396557e-108  1.556329e-01  8.480347e-03 1.098440e-108
##
##
## $results_mult_reg
## $results_mult_reg$point_estimates
##     alpha_0     alpha_1    alpha_XY     alpha_2
## -0.01822490  0.27671667  0.13346492 -0.01232202
##
## $results_mult_reg$SE_estimates
##    alpha_0    alpha_1   alpha_XY    alpha_2
## 0.03687934 0.02978589 0.05247700 0.02960048
##
## $results_mult_reg$pvalues
##      alpha_0      alpha_1     alpha_XY      alpha_2
## 6.212906e-01 9.406853e-20 1.113124e-02 6.772964e-01
##
##
## $results_res_reg
## $results_res_reg$point_estimates
##     alpha_0     alpha_1     alpha_2     alpha_3    alpha_XY
##  0.03444987  0.28261304 -0.01268054 -0.05279386  0.13264790
##
## $results_res_reg$SE_estimates
##    alpha_0    alpha_1    alpha_2    alpha_3   alpha_XY
## 0.03059805 0.02977685 0.02968121 0.03680861 0.05226472
##
## $results_res_reg$pvalues
```

```
##      alpha_0      alpha_1      alpha_2      alpha_3     alpha_XY
## 2.604853e-01 1.628744e-20 6.693082e-01 1.518056e-01 1.129989e-02
##
##
## $results_sem
## $results_sem$point_estimates
##     alpha_1     alpha_3     alpha_4     alpha_6    alpha_XY
## -0.01411334  0.13713065 -0.04066894  0.27722407  0.13356893
##
## $results_sem$SE_estimates
##    alpha_1    alpha_3    alpha_4    alpha_6   alpha_XY
## 0.05593721 0.05554422 0.03139961 0.02970394 0.05237590
##
## $results_sem$pvalues
##      alpha_1      alpha_3      alpha_4      alpha_6     alpha_XY
## 8.008039e-01 1.355460e-02 1.952495e-01 1.030064e-20 1.076616e-02
##
##
## attr(,"class")
## [1] "ciee"

# Wrapper of ciee function to analyze multiple SNVs sequentially
maf <- 0.2
n <- 1000
dat <- generate_data(n = n, maf = maf)
datX <- data.frame(X = dat$X)
names(datX)[1] <- "X1"
for(i in 2:10){
    X <- rbinom(n, size = 2, prob = maf)
    datX$X <- X
    names(datX)[i] <- paste("X", i, sep="")
}
results_ciee_loop <- ciee_loop(setting = "GLM", Y = dat$Y, X = datX,
                               K = dat$K, L = dat$L)
results_ciee_loop

## $results_ee
## $results_ee$point_estimates
##           X1           X2           X3           X4           X5
##  0.001294153  0.006898206 -0.021326542  0.007787436  0.002209479
##           X6           X7           X8           X9          X10
##  0.019024904  0.024567674  0.090631472 -0.069561087  0.016969094
##
## $results_ee$SE_estimates
```

```
##          X1          X2          X3          X4          X5          X6
## 0.05284822 0.05249032 0.05825832 0.05431078 0.05422003 0.05004800
##          X7          X8          X9         X10
## 0.05248844 0.05156188 0.05333335 0.05667113
##
## $results_ee$wald_test_stat
##           X1           X2           X3           X4           X5
##  0.02448810  0.13141862 -0.36606860  0.14338657  0.04075024
##           X6           X7           X8           X9          X10
##  0.38013314  0.46805881  1.75772234 -1.30427003  0.29943103
##
## $results_ee$pvalues
##          X1          X2          X3          X4          X5          X6
## 0.98046327 0.89544416 0.71431389 0.88598489 0.96749501 0.70384659
##          X7          X8          X9         X10
## 0.63974254 0.07879476 0.19214153 0.76461119
##
##
## $results_mult_reg
## $results_mult_reg$point_estimates
##            X1           X2           X3           X4           X5
##  0.001288740  0.006962460 -0.021669079  0.007790108  0.002254990
##            X6           X7           X8           X9          X10
##  0.018939691  0.024558364  0.090638752 -0.070201044  0.017164744
##
## $results_mult_reg$SE_estimates
##          X1          X2          X3          X4          X5          X6
## 0.05209189 0.05347009 0.05475888 0.05307842 0.05304104 0.05215047
##          X7          X8          X9         X10
## 0.05442412 0.05464559 0.05475659 0.05763805
##
## $results_mult_reg$pvalues
##          X1          X2          X3          X4          X5          X6
## 0.98026751 0.89642480 0.69239776 0.88334641 0.96609743 0.71655204
##          X7          X8          X9         X10
## 0.65191450 0.09749807 0.20012111 0.76591616
##
##
## $results_res_reg
## $results_res_reg$point_estimates
##            X1           X2           X3           X4           X5
##  0.001270285  0.006951005 -0.021511171  0.007777435  0.002252196
##            X6           X7           X8           X9          X10
```

```
##  0.018865254  0.024532738  0.090519496 -0.069738927  0.017128832
##
## $results_res_reg$SE_estimates
##         X1         X2         X3         X4         X5         X6
## 0.05166572 0.05337253 0.05450433 0.05298206 0.05295503 0.05199572
##         X7         X8         X9        X10
## 0.05434119 0.05455498 0.05452165 0.05752001
##
## $results_res_reg$pvalues
##         X1         X2         X3         X4         X5         X6
## 0.98038961 0.89640624 0.69317148 0.88332447 0.96608443 0.71681381
##         X7         X8         X9        X10
## 0.65175794 0.09738278 0.20115730 0.76592610
##
##
## $results_sem
## $results_sem$point_estimates
##           X1           X2           X3           X4           X5
##  0.001280605  0.006893870 -0.021317838  0.007792080  0.002205729
##           X6           X7           X8           X9          X10
##  0.019020168  0.024571386  0.090635872 -0.069570486  0.016969045
##
## $results_sem$SE_estimates
##         X1         X2         X3         X4         X5         X6
## 0.05198753 0.05335219 0.05450744 0.05297224 0.05292888 0.05202060
##         X7         X8         X9        X10
## 0.05431481 0.05453628 0.05449757 0.05746261
##
## $results_sem$pvalues
##         X1         X2         X3         X4         X5         X6
## 0.98034776 0.89718802 0.69572361 0.88305513 0.96675902 0.71464299
##         X7         X8         X9        X10
## 0.65098926 0.09652534 0.20175073 0.76776026
##
##
## attr(,"class")
## [1] "ciee"

# Summary functions of the CIEE output
summary(results_ciee)

## [1] "Results based on estimating equations."
##                point_estimates SE_estimates wald_test_stat    pvalues
## CIEE_alpha_0       -0.01822490   0.03751627     -0.4857865  6.271e-01
```

```
## CIEE_alpha_1        0.27671667   0.03147702      8.7910704  1.481e-18
## CIEE_alpha_2        0.13346492   0.05078096      2.6282473  8.583e-03
## CIEE_alpha_3       -0.01232202   0.02879671     -0.4278968  6.687e-01
## CIEE_sigma_1_sq     0.92029662   0.04157306     22.1368523  1.40e-108
## CIEE_alpha_4       -0.05318825   0.03745880     -1.4199133  1.556e-01
## CIEE_alpha_XY       0.13363882   0.05076842      2.6323219  8.480e-03
## CIEE_sigma_2_sq     0.92045700   0.04155998     22.1476747  1.10e-108
## [1] "Results based on traditional multiple regression."
##             point_estimates SE_estimates      pvalues
## MR_alpha_0      -0.01822490   0.03687934 6.212906e-01
## MR_alpha_1       0.27671667   0.02978589 9.406853e-20
## MR_alpha_XY      0.13346492   0.05247700 1.113124e-02
## MR_alpha_2      -0.01232202   0.02960048 6.772964e-01
## [1] "Results based on traditional regression of residuals."
##             point_estimates SE_estimates      pvalues
## RR_alpha_0       0.03444987   0.03059805 2.604853e-01
## RR_alpha_1       0.28261304   0.02977685 1.628744e-20
## RR_alpha_2      -0.01268054   0.02968121 6.693082e-01
## RR_alpha_3      -0.05279386   0.03680861 1.518056e-01
## RR_alpha_XY      0.13264790   0.05226472 1.129989e-02
## [1] "Results based on structural equation modeling."
##              point_estimates SE_estimates      pvalues
## SEM_alpha_1      -0.01411334   0.05593721 8.008039e-01
## SEM_alpha_3       0.13713065   0.05554422 1.355460e-02
## SEM_alpha_4      -0.04066894   0.03139961 1.952495e-01
## SEM_alpha_6       0.27722407   0.02970394 1.030064e-20
## SEM_alpha_XY      0.13356893   0.05237590 1.076616e-02

summary(results_ciee_loop)

## [1] "Results based on estimating equations."
##          point_estimates SE_estimates wald_test_stat    pvalues
## CIEE_X1      0.001294153   0.05284822     0.02448810 0.98046327
## CIEE_X2      0.006898206   0.05249032     0.13141862 0.89544416
## CIEE_X3     -0.021326542   0.05825832    -0.36606860 0.71431389
## CIEE_X4      0.007787436   0.05431078     0.14338657 0.88598489
## CIEE_X5      0.002209479   0.05422003     0.04075024 0.96749501
## CIEE_X6      0.019024904   0.05004800     0.38013314 0.70384659
## CIEE_X7      0.024567674   0.05248844     0.46805881 0.63974254
## CIEE_X8      0.090631472   0.05156188     1.75772234 0.07879476
## CIEE_X9     -0.069561087   0.05333335    -1.30427003 0.19214153
## CIEE_X10     0.016969094   0.05667113     0.29943103 0.76461119
## [1] "Results based on traditional multiple regression."
##        point_estimates SE_estimates    pvalues
```

```
## MR_X1      0.001288740   0.05209189 0.98026751
## MR_X2      0.006962460   0.05347009 0.89642480
## MR_X3     -0.021669079   0.05475888 0.69239776
## MR_X4      0.007790108   0.05307842 0.88334641
## MR_X5      0.002254990   0.05304104 0.96609743
## MR_X6      0.018939691   0.05215047 0.71655204
## MR_X7      0.024558364   0.05442412 0.65191450
## MR_X8      0.090638752   0.05464559 0.09749807
## MR_X9     -0.070201044   0.05475659 0.20012111
## MR_X10     0.017164744   0.05763805 0.76591616
## [1] "Results based on traditional regression of residuals."
##        point_estimates SE_estimates    pvalues
## RR_X1      0.001270285   0.05166572 0.98038961
## RR_X2      0.006951005   0.05337253 0.89640624
## RR_X3     -0.021511171   0.05450433 0.69317148
## RR_X4      0.007777435   0.05298206 0.88332447
## RR_X5      0.002252196   0.05295503 0.96608443
## RR_X6      0.018865254   0.05199572 0.71681381
## RR_X7      0.024532738   0.05434119 0.65175794
## RR_X8      0.090519496   0.05455498 0.09738278
## RR_X9     -0.069738927   0.05452165 0.20115730
## RR_X10     0.017128832   0.05752001 0.76592610
## [1] "Results based on structural equation modeling."
##         point_estimates SE_estimates    pvalues
## SEM_X1      0.001280605   0.05198753 0.98034776
## SEM_X2      0.006893870   0.05335219 0.89718802
## SEM_X3     -0.021317838   0.05450744 0.69572361
## SEM_X4      0.007792080   0.05297224 0.88305513
## SEM_X5      0.002205729   0.05292888 0.96675902
## SEM_X6      0.019020168   0.05202060 0.71464299
## SEM_X7      0.024571386   0.05431481 0.65098926
## SEM_X8      0.090635872   0.05453628 0.09652534
## SEM_X9     -0.069570486   0.05449757 0.20175073
## SEM_X10     0.016969045   0.05746261 0.76776026
```